VISIT…

LANZAROTE Caliente.COM

许国璋
王宗炎 主编

现代语言学丛书

应用语言学

刘涌泉 乔毅 编著

上海外语教育出版社

现代语言学丛书

应用语言学

刘涌泉　乔　毅　编著

戚雨村　审订

上海外语教育出版社

文忞连

1998. 3. 8

现代语言学丛书

应 用 语 言 学

刘涌泉　乔　毅　编著

戚雨村　审订

上海外语教育出版社出版发行

（上海外国语大学内）

商务印书馆上海印刷厂印刷

新华书店上海发行所经销

开本 850×1168　1/32　7.875 印张　203 千字

1991 年 8 月第 1 版　1996 年 10 月第 2 次印刷

印数：5 001－7 000 册

ISBN 7－81009－564－1

H·316　定价：9.00 元

《现代语言学丛书》
编委名单

主　编　许国璋
　　　　　王宗炎

D568/08

编　委　(按姓氏笔划为序)
王宗炎(中山大学)
王德春(上海外国语学院)
许国璋(北京外国语学院)
伍铁平(北京师范大学)
张日昇(香港中文大学)
赵世开(中国社会科学院语言研究所)
桂灿昆(广州外国语学院)
桂诗春(广州外国语学院)
戚雨村(上海外国语学院)
缪锦安(香港大学)

总　序

为什么出版《现代语言学丛书》?

因为我们感到，中国现代化包括许多方面的工作，其中之一是语言学研究的现代化。我们希望这一套丛书的出版，会有助于这一工作的开展。

近几十年来，国外语言学的研究进展很快。一方面，关于语言的内部结构，出现了各种理论和模式；另一方面，从各种不同的学科去研究语言，产生了诸如人类语言学、社会语言学、心理语言学、神经语言学、计算语言学等多科性研究。了解和介绍这两方面的理论、模式、实验和数据，供我国语言研究者参考，从而为语言学研究的现代化出一点力，这是我们的希望。

要做到语言学研究的现代化是不容易的。首先要对国外新的语言学理论加以分析和比较，作出我们自己的判断；更重要的是要结合汉语的研究加以验证，写出结合中国实际的论著。我们这里先做第一步工作。

中国语言学史上，不乏利用外国的语言理论，为汉语研究开辟新路的例子。郑樵说："初韵之学，起自西域。"马建忠以拉丁文法为范式，写出了《马氏文通》。赵元任、罗常培等前辈先生运用描写语言学的方法，为我国方言调查做出了典范。近时汉语语法学家利用国外语言学的研究方法，使语法现象的分类和范畴的描写更有理据，更为精确。先行者研究外国语言理论的态度，永远是值得我们学习的。

作为第一步，我们打算出版15至20种书。以普及为主，逐步

提高；以引进为主，同时注意结合我国的实际。我们希望和国内语言学界同志共同努力，填补我国语言学科中的一些空白点。

我们心目中的读者，是高等学校中文、外文和其他文史专业的师生，翻译界、新闻出版界人士，中学语文教师，以及一般语文工作者和爱好者。我们将力求用明白易懂的语言介绍新的学说和理论。

我们将注意国外新出的语言学文献，为中国的语言学的现代化尽快提供信息。我们的力量还很薄弱，我们要努力去做，并热诚希望国内语言学者和语文工作者给予指导、批评和支持。

《现代语言学丛书》编委会

一九八二年十一月

前　言

应用语言学这个名称，是波兰语言学家J·N·博杜恩·德·库尔特内在19世纪70年代提出来的。在他的著作中，应用语言学相对于理论语言学而言，其对象是把理论语言学的知识应用于解决其他科学领域的问题。尽管对应用语言学有不同理解，但有一点是明确的：这门学科不是指语言理论方面的研究，而是指语言应用方面的研究。应用语言学可以定义为研究语言在各个领域中实际应用的学科。

语言理论和语言应用有密切的关系。甚至可以说，语言应用是促使语言理论研究的动力。吕叔湘先生就曾指出，语言学最初的建立，就是为读古书和学作文服务的。但只是在本世纪40—50年代，应用语言学才发展成为一门独立的学科，近些年来，更取得重大的进展。也正因为如此，它的研究对象、涉及范围以及学科体系都还没有定型。

应用语言学一般有广狭两义。广义的应用语言学泛指把语言学知识应用于其他科学领域，狭义的应用语言学专指语言教学，特别是第二语言教学和外语教学。根据近年来的发展趋势看，似有必要区分为一般应用语言学和机器应用语言学，前者处理面向人的一些问题，后者处理面向电子计算机的一些问题。一般应用语言学包括以下一些方面：语言教学、标准语的建立和规范化、文字的创制和改革、辞书编纂、翻译、标音和转写等；机器应用语言学研究得比较多的问题有：实验语音学、机器翻译、情报检索、语文信息处理、自然语言理解、言语统计等。

应用语言学以语言学为基础，但不仅限于语言理论和语言描写的知识。比如说，语言教学就要吸收并应用语言学、教育学、心理学、教育心理学、心理语言学等学科的知识，自然语言理解则与语言学、心理学、逻辑学、声学、数学和计算机科学等学科的知识有关。

本书的内容既涉及一般应用语言学，也涉及机器应用语言学，而只就其中的某些问题作些阐述和说明。因此，这不是一本全面、系统地论述应用语言学的专著，而只是这门学科的一本入门读物。

编　者

一九八九年十二月

目　　录

1. 应用语言学概说 …………………………………… 1
2. 语种识别 ……………………………………………12
3. 标音、转写和译音 ………………………………44
4. 术语 …………………………………………………71
5. 机器翻译 ……………………………………………91
6. 情报检索 …………………………………………167
7. 言语统计 …………………………………………179
8. 计算机辅助语言教学 ……………………………186
9. 语音信息处理 ……………………………………200
10. 中文信息处理 ……………………………………216

1. 应用语言学概说

在科学技术大发展的年代，语言学也得到了很大的发展。这种发展表现在许多方面，例如，数理语言学作为语言学的一个分支已站稳了脚跟；从结构语言学只注重形式研究业已发展为注重语义研究；继转换生成语法学派之后，最近又兴起了语用学(pragmatics)和篇章语言学（text linguistics)。值得强调指出的是，应用语言学的长足进展，则是这种发展的一个最显著的特征。

1.1 什么是应用语言学 应用语言学（applied linguistics)是语言学知识在各个领域中各种不同应用的总称，它研究语言如何得到最佳利用的问题。狭义地讲，应用语言学主要是指语言教学。广义地讲，它所涉及的领域则是很多的。本章拟对它作一个总括的描述。

1.1.1 特征和意义 应用语言学是语言学的一个重要方面。它与理论语言学不同，着重解决现实当中的实际问题，一般不接触语言的历史状态；而理论语言学则着重探索理论问题，在总结规律时常涉及语言的历史状态。当然，两者之间的关系是密不可分的：理论要以应用(实践)为基础，应用又以理论为指导。可以说，语言理论是应用语言学的基本武器，而应用语言学则是鉴定各种理论的实验场。

语言是人类最重要的交际工具。语言的发展总是与社会的发展密切相关的。人们在运用语言的过程中又总是不断改善和扩大

它的交际职能。这就是应用语言学产生和发展的原因。最初的语言，只有有声语言的形式。后来出现了文字，产生了书面语言，语言的应用扩大了。从时间上说，它能流传到后代，从空间上说，它能传递到远方。后来又出现了印刷机，语言的应用得到进一步扩大。不仅能流传到后代和远方，而且能比较快速地大量印刷，广为传播。录音机和电话的出现，又是一次飞跃，从此，语言不但可以保留其书面形式，而且也可以保留其语音形式，不但能以书面形式，而且能以语音形式迅速传递到远方。电视、传真、复印、录象、激光照排、卫星通讯、模式识别、各种语言分析合成仪器、电子计算机等先进技术和设备的出现，更把语言的应用提高到一个新水平。在“计算机文化”到来的社会里，语言已不仅是人与人之间的交际工具，而且是人机对话的基础。必须指出，语言应用范围每扩大一步，都会使语言的交际职能得到明显的提高。同时也会出现一些新问题。为了解决这些问题，常常又必须采用一些新的方法和新的工具，而新方法和新工具的应用又反过来促进应用语言学以及整个语言学的发展。

1.1.2 历史和发展　语言研究的应用方面和理论方面最初没有划分开来，这是因为语言学不够发达，还没有多少必要来作这种区分。19世纪初，理论方面的研究和应用方面的研究开始分化，例如，作为应用语言学一个分支的语言教学同当时着重探讨历史问题的语言学分了手。19世纪下半叶，博杜恩·德·库尔特内（J. Baudouin de Courtenay, 1845—1929）就提出了“应用语言学”这个概念，但没有得到广泛的注意。20世纪以来，语言科学得到进一步的发展。特别是近三四十年，各方面向语言科学提出了一系列与科学技术、国民经济的发展有密切关系的新任务，应用范围得到了空前的扩大，这才使得语言应用方面的研究和理论方面的研究明确地区分开来，“应用语言学”这个名词开始广泛运用。

语言学由于应用范围的扩大，同其他学科的联系也增加了。

过去语言学只同文学、人类学、历史学、地理学、考古学、心理学、哲学和文化史等有联系。而今，它同数学、信息论、控制论、物理学、电子学、医学、符号学、情报学、计算机科学、通讯技术、自动化技术等发生了密切联系。语言学的领域如此扩展，无疑地带来了新的分工和一些部门的分化。有的人进行历史比较，有的人钻研理论，有的人去解决实际问题，还有一些人则去研究一些最新的、对国民经济有重大作用、而且与尖端学科相联系的边缘问题。所有这一切，都促成了应用语言学与理论语言学的分化。

1.2 任务和范围 应用语言学是一个较大的概念，可以说，同语言学应用有关的问题都属于应用语言学的研究范围。它同其他应用学科一样，是以直接满足社会需要为目标的，因此，它的研究范围由实践的需要来决定。根据近年来的发展趋势看，有必要区分为一般应用语言学和机器应用语言学。前者处理面向人的一些老问题，后者处理面向机器的新问题。

1.2.1 一般应用语言学

1.2.1.1 语言教学研究　语言教学是应用语言学中最古老的一个分支，但成为一个独立学科是近几十年的事。编辑高质量的教材和参考书，研究切合实际的教学方法，一直是语言教学研究中的重大课题。随着文化教育事业的发展，除一般的语言教学外，还开展了为不同目的和不同对象服务的第二语言教学、科技外语教学、双语教学、聋哑盲教学，等等。

在具有悠久文化历史的国度里，语言教学的任务更加繁重，以中国来说，古汉语文献汗牛充栋，古典文学作品极其丰富。如何在语文教学中培养学生既掌握好现代汉语的基础知识和写作能力，又获得一定的古汉语知识和欣赏古典文学作品的能力，这是一个特殊课题。

对于多民族国家来说，除搞好各民族本族语言教学外，如何搞好第二语言(母语以外)教学也是一项重大任务。

在当今世界，一个国家，尤其是发展中国家，如何把外语教学搞好，使人们获得掌握先进科学技术的工具，是摆在语言学家和外语教师面前的重要职责。

相反地，如何让外国人学好某一种语言（如汉语），以便世界各地的朋友共享该语言创造的文化资源，同样是一个重要课题。

近年来，随着科学技术的发展，语言教学也在经历一个现代化的过程。语言实验室、计算机辅助语言教学应运而生。

1.2.1.2 标准语 标准语的建立和规范化，文字的创制和改革也是应用语言学研究的范围。这里面既有政策问题，又有技术问题。建立通用于各方言区的标准语，对于一个民族或一个国家的统一和发展是极端重要的。在方言分歧的情况下，如何选择好这种标准语的基础方言，又如何确定它的标准音，这是应用语言学所要解决的问题。为无文字的语言创制文字时，基础方言和标准音更是重要的依据。由于这种语言多半是从未研究过的，因而语言学家需要进行周密细致的语言调查，对比各方言点的音位系统，并在此基础上结合语言使用情况及政治经济文化方面的因素对基础方言和标准音加以确定。文字改革包括文字系统（字母表、正字法和标点符号）的部分改进和文字系统的彻底更换。前者如苏联1918年俄文正字法改革，后者如1928年土耳其的文字改革（以拉丁字母取代了原有的阿拉伯字母）。

标准语的建立只是语言规范化的开始。为了确定语音、语法、词汇规范，需要编出相应的正音词典、规范语法和各种类型的词典。同时，还要采取各种措施提高人们的语言修养。

1.2.1.3 辞书编纂 辞书编纂是应用语言学中同实际联系最密切的学科之一，同时也是反映语言研究水平的一个重要侧面。比如，对词类问题研究不够，就不好在词典中为每个词标记词类。又如，某个字古音不详，在汉语历史大辞典中只好付诸阙如。莫怪人们常说，词典编纂是记录之学。

词汇是语言中变化最快的部分，新词新义不断涌现。及时地、

准确地把这些新词新义收在词典中，指导人们如何运用，这是辞书对语言规范化最有效的影响。

为了满足人们多方面的需要，辞书品种日益增多。除一般性字典、词典外，还编纂了同音字典、同音词典、构词词典、正字正词词典、同义反义词典、类语词典、词源词典、成语词典、俚语词典、谚语词典、典故词典、惯用语词典、歇后语词典、人名地名词典、各种方言词典、各种术语词典、缩略语词典、大百科全书，最近又出版了很多频率词典和逆序词典。辞书业的兴盛反映了文化教育事业的发达。

1.2.1.4 翻译 翻译是文化交流的桥梁。直译和意译问题从古至今一直是一个有争论的问题。“信、达、雅”更是近年来争论的焦点。这些问题虽说带有一定的理论性，但也同翻译实践密切相关。至于传授翻译技巧，如何使译文达到与原文神似的程度，更是各种翻译教科书和翻译手册所要讨论的主要内容。

1.2.1.5 术语、人名和地名翻译 术语、标音和转写、人地名研究这些问题与翻译有密切联系。人名、地名的翻译一般都是根据“名从主人”的原则加以音译。但是，在各种不同的语言文字之间如何音译，采取标音法还是转写法，也是尚未完全解决的问题。不过，目前的趋势是转写法占上风。在使用汉字的情况下，问题更加复杂，一般的标音和转写原则不适用，因而制定出各种对译不同语言的汉字译音表。使用汉字译音，带来了译名的驳杂现象（如 Reagen：“里根”、“雷根”，New Haven：“纽黑文”、“新哈芬”）。为了消除这种驳杂现象，并为了更好地贯彻“名从主人”的原则，不少人主张利用汉语拼音照抄拉丁字母的人名地名，并用它转写斯拉夫等字母书写的人名地名。术语是科技语言最基本的载信单元。术语混乱必然要影响国内外学术交流和科学技术的进步。因而各个国家和国际学术组织（如国际标准化组织 ISO）对术语的标准化十分重视。随着科学技术的迅猛发展，术语大量涌现，译不胜译，因而术语国际化的倾向日益发展。如果音译，又

发生怎样音译的问题，是标音还是转写，或是采取某种折衷的方法。不过有一点是明确的：术语不同于人名、地名，移植进来之后便变成该语言词汇的一部分，这就是说，它不仅要化在其中，而且要生根开花。

1.2.1.6 其他 除了上面这些课题外，一般应用语言学还涉及下列内容：(1)言语矫正学，协助医生治疗各种先天的和后天的言语病症(如颚裂、口吃、倒仓、塌中等等)，使患者的语言能力得到恢复。(2)舞台语言研究，目的是领导演员如何进行发声训练，获得艺术语言的基本功。(3)建立国际辅助语，目的是探讨消除语言障碍的途径。(4)制定速记系统，以便于人们用简单的符号迅速记录语言。

1.2.2 机器应用语言学 电子计算机的出现，引起了科学技术的巨大变化，同时也为语言学开辟了新的发展途径。计算机一方面充当语言学工作者的得力助手，帮助语言学工作者对语言素材进行分类、模拟、分析和转换，另一方面又对语言学提出了一系列新要求，要求武装它的“头脑”以发展它的智力（如赋于它检索能力、翻译能力），给它增添“翅膀”以赋于它听觉（识别口语）、更强的视觉(识别文字，包括语种识别)，说话能力(言语合成)和听写能力(语音打字)。为了处理语言应用方面的这些问题，而且主要是跟机器打交道的一些新问题，工程语言学、机器语言学、计算语言学等应运而生。所有这些便形成了机器应用语言学的内容，其特点是研究如何利用电子计算机等先进工具来处理自然语言。目前研究得比较多的课题有以下几个。

1.2.2.1 实验语音学 20世纪初已萌芽，但真正成为一门实验性学科是近二三十年的事。最初，研究内容的发音器官和发音方法的生理实验为主。实验工具多借助医疗器械。后来发展以声学实验为主，为通讯工程提供各种参数。语图仪、X光电影、动态颚位记录装置等分析仪器的出现，使语音研究从静态分析过渡到动态分析。电子计算机的应用，更使语音实验研究发展到一个新阶

段。由于计算机有快速处理大量数据的能力，成段语言的分析研究才得以进行。于是，语音实验从音素音节分析扩展到成句成章分析，同时超音段特征成了重要研究对象。这些都给言语的分析和合成创造了有利条件。由此可见，实验语音学作为一门边缘学科在给电子计算机增添听说能力方面将起决定性作用。

1.2.2.2 机器翻译 电子计算机和语言的最早结合开始于机器翻译。作为第一次结合，这个选择可以说又成功又不成功。说成功，一是机器翻译好比一颗火种，点燃了计算机非数值应用的革命，二是它的确成了一个无比广阔的实验场，许多语言学理论和方法以及许多技术成果都是在它的基础上或启发下产生或解决的。说不成功，是由于机器翻译是一种比较高级的人工智能，难度大，不仅要求对一种自然语言有充分理解，而且要求把它等价地转换成另一种自然语言，因而至今尚未能真正或广泛地付诸应用，没有起到示范作用。

机器翻译经历了一个马鞍形的发展过程，从高潮到低潮，现在又进入一个日趋繁荣的新阶段，并成了第五代计算机主要内容之一。实现机器翻译的技术条件目前已基本具备，存在的主要问题是语言研究跟不上。不过，现在世界上已有十多个机器翻译系统在初步应用。

1.2.2.3 情报检索 科技情报的数量每8—10年就翻一番，在大量的资料中查找某些需要的东西，犹如大海捞针。手工检索和机械检索已进行多年，成效有限。只是由于采用了电子计算机，情报检索才出现崭新的面貌，整个情报工作才得以进入“情报—计算机—电讯”三位一体的新时期。

情报检索系统中的关键问题是情报检索语言的建立。这种语言应具备能精确表达文献主题和提供主题所需的词汇语法手段，不应产生歧义，不受用户主观因素影响，并且便于用程序运算方式进行检索。为了提高情报检索的效能，使其具有较高的查全率和查准率，在词汇方面如何消除术语的同义性和多义性，在语法

方面如何求得既经济又足够的语法手段来表达必要的语法关系，这些都需要进行认真的研究。对于中文情报检索来说，还有一个特殊问题，即词的切分问题需要解决。

1.2.2.4 汉字信息处理 随着计算机非数值应用范围的不断扩大，汉字信息处理问题越来越显得尖锐。这个问题不很好的解决，势必影响中文信息处理各项工作的开展。汉字字形繁复，字数庞杂，而且存在大量一音多字、一字多音现象。这给编码输入带来很大麻烦。目前虽然已有一些方案初步解决了输入输出问题，但如何求得比较理想的方案，使编码做到简单易学，操作方便，输入迅速，没有重码，而且易于进一步处理，这还是一个尚待解决的问题。为了合理解决这个问题，有必要对汉字进行多方面的（数理的、语言学的、工程心理学的）分析研究。从语言学方面说，加强汉字的基础研究不仅对汉字信息处理是必要的，而且也是语言文字学的迫切任务。如书面语的统计研究（词、字、部件、笔划等）和口语的统计研究（多音节词、音节、音素、声调等）不仅有利于编码方案的设计，而且也有利于输入输出装置的设制，例如键盘的设计和字库的安排。汉字信息处理是中文信息处理研究中最关键的问题，目前来说，不解决这个问题，任何中文信息处理系统都不能推广应用。

1.2.2.5 自然语言 自然语言理解是人工智能的核心，因为人的各种智能活动最终都要以语言及其书面代表——文字来表达。自然语言理解中的困难问题，同机器翻译等机器语言学部门存在的问题一样，主要是区别各种同形多义现象，以及解决省略、替代等问题。

同形现象分布在各个平面上，以汉语为例，词的平面上，如“打”这个词可以有“打人”、“打哪儿来”、“打篮球”、“打二两酒”、“打毛衣”、“打铺盖卷儿”、“打电报”、“一打铅笔”等意义。句子平面上，如“反对的是少数人”，其中“反对的”，可以是“反对者”（施事），也可以是“被反对者”（受事）。

以上这些同形现象，除少数可以靠词本身加以排除外，一般都要靠句内上下文通过语义结构手段来确定，个别的还需要依靠更大范围的上下文，或语用学知识。

1.2.2.6 言语统计　由于应用了电子计算机，言语统计这门老学科获得了新生。过去用手工方式进行统计，不仅统计量不可能很大，而且有些数据，如语言单位的出现概率，也无法提供。近年来，利用电子计算机进行统计，既快又准，统计量不受限制，而且能提供多种参数，因而促进了统计语言学的迅速发展。

进行言语统计，目的在于根据量的描述给出质的评价，即依靠定量分析得出定性分析。例如，人们对词进行频率统计，统计结果告诉我们某些词出现几千、几万次，而某些词只出现三五次，甚至一次。显然，我们可以给出这样的质的评价：前一类词是常用词，后一类词是非常用词。这些数据无疑对语言教学等工作是极为重要的参考。随着科学技术和文化教育的发展，言语统计范围不断扩大。例如，为了解决工程中提高清晰度压缩言语信号的问题，进行了元音、辅音以及韵律特征等大量语音参数的统计；为了对语法进行精确描写，统计了语法形式、单位和模式的用法；为了找到各种功能风格以及不同作家的风格特点，开展了计算风格学研究；为了确定标准汉字表和计算机内部标准码，进行了大量汉字字频统计。

进行言语统计，需要各种素材，这些素材的合理组织，便构成了语料库。语料库编排得合理与否，直接影响到语料效果的发挥。编排得好，提供的副产品就多，例如可以利用所存的语料编制逐词索引（concordance）和建立试题库，或进行各种对比研究。

1.2.2.7 少数民族语文的信息处理　中国一些少数民族至今仍使用古老的文字，书写印刷非常不便。近年来，继中文信息处理蓬勃发展之后，民族语文信息处理的研究也开展起来了。目前，已经进行语言信息处理研究的民族有蒙族、藏族、维吾尔族、哈萨

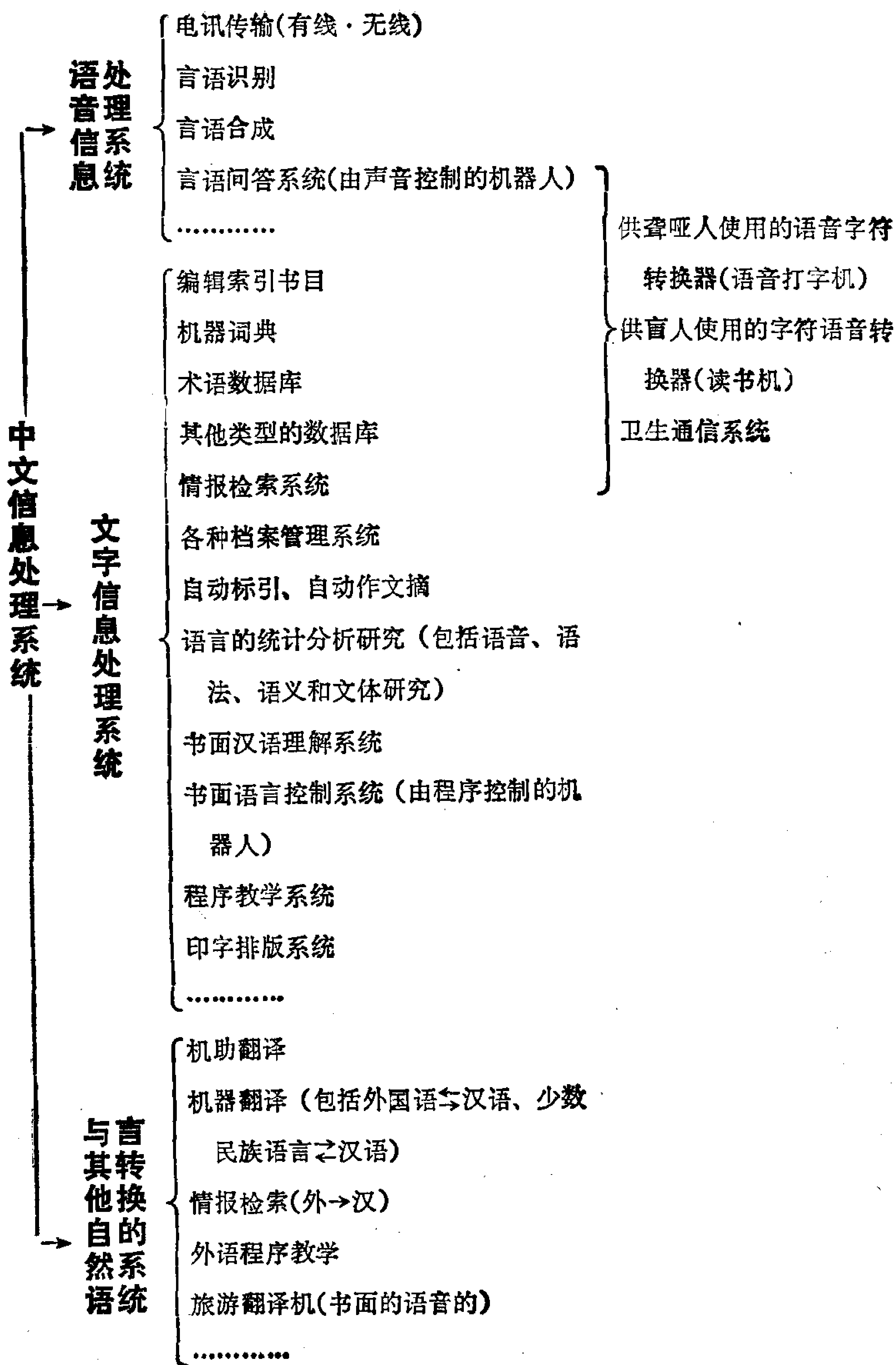
克族、柯尔克孜族、壮族、彝族和朝鲜族，其他一些民族语言的信息处理也在酝酿之中。无疑地，这将大大提高这些民族的语言文字应用水平，促进他们的文化教育事业和整个社会主义建设事业的发展。

1.3 发展前景 语言应用是不断向前发展的，文化的进步，新技术的涌现，使语言应用的范围日益扩大，使语言应用的潜力日益发挥，使语言应用的效果日益完善。目前大家都在谈论世界正在进入信息化社会，从语言学的观点看，所谓信息化社会，其主要的特征之一就是利用电子计算机等先进工具对语言文字信息进行各种处理(包括语言文字信息的储存、分类、统计、检索、转换、传输、控制和模拟等)，目的在于建立现代化语言信息系统，使语言文字得到最佳利用，使凝聚在语言文字中的知识信息发挥最大效能。

应用语言学的发展前途无量，随着新技术的不断涌现，全国性、地区性，甚至国际性的现代化语言信息系统将逐步建成。它是一个极其复杂的巨系统。在这个巨系统下有文字信息处理和语音信息处理两个大系统，而每个大系统之下又有一些分系统和小系统。以中文为例，整个系统见下页图示。

当然，建立这样的系统并不是一件轻而易举的事。它要求语言学、计算机科学、通讯技术、自动化技术等方面的专家共同努力才能完成。对于语言学家来说，任务更加繁重，因为说到底，上述系统的各个部分毕竟主要是语言文字问题。因此，必须当作语言规划 (language planning) 中的主要课题来抓。

由于应用语言学涉及的课题很多，本书以下各章难以面面俱到。我们讨论的重点只是那些与我们近年来的研究工作关系较为密切的课题。


中文信息处理系统
语音信息处理系统
电讯传输(有线·无线)
言语识别
言语合成
言语问答系统(由声音控制的机器人)
…………
供聋哑人使用的语音字符转换器(语音打字机)
供盲人使用的字符语音转换器(读书机)
卫生通信系统
文字信息处理系统
编辑索引书目
机器词典
术语数据库
其他类型的数据库
情报检索系统
各种档案管理系统
自动标引、自动作文摘
语言的统计分析研究(包括语音、语法、语义和文体研究)
书面汉语理解系统
书面语言控制系统(由程序控制的机器人)
程序教学系统
印字排版系统
…………
言语与其他自然语转换的系统
机助翻译
机器翻译(包括外国语⇆汉语、少数民族语言⇄汉语)
情报检索(外→汉)
外语程序教学
旅游翻译机(书面的语音的)
…………

2. 语种识别

2.1 语种识别 语种识别是应用语言学中一个不常受到重视，但却很有意义的课题。它是图书、情报工作者和科学技术人员经常遇到的问题之一。这些人虽然不一定要掌握多种外语，但学会准确地鉴别各种书刊的语种却是必要的和可能的。情报学与应用语言学有着十分密切的联系，语言学实际上是情报学的一个有机组成部分，是它的基础之一。

在各种外文图书刊物中，有许多语言种类。有英、俄、日、德、法这五大语种，还有许多其他语种，例如：意大利语、瑞典语、波兰语、捷克语、匈牙利语、罗马尼亚语……等等。这样便会遇到按语种将书刊进行分类的问题，也就产生了语种识别问题。当然，识别语种和掌握语言是不同的概念。但是，为了进行语种识别必须依照一套特定的算法，建立识别规则。

首先让我们规定一下语种识别的范围。世界上的语言共有三四千种，但只有两百多种应用得较为广泛。其中，占世界总人口95%以上的人所用的语言还不到一百种。作为语言的书面表达形式的文字，在当今世界上约有六七十种最为常用。在各类文字当中，有的使用拉丁字母，有的使用斯拉夫字母，有的使用阿拉伯字母，有的使用梵文字母，有的使用各自的民族字母。一般说来，除专用于某语种的民族字母之外，我们都无法根据字母类型来进行语种识别。在最常用的六七十种文字当中，使用拉丁字母的语言就有二三十种，使用斯拉夫字母的语言也有十来种。本章准备分别讨论一下使用拉丁字母和斯拉夫字母的语种识别问题。下面

先谈使用拉丁字母的 27 种语言的识别问题。

2.2 拉丁字母的语种识别 既然有这么多语言都使用拉丁字母，究竟怎样识别它们呢？我们认为，要解决这一问题，就必须找出各语种言在文字形式方面的区别特征和相互间的差异。具体地说，语种识别的关键性的技巧就在于考察各种文字究竟采用哪些区别符号，这些区别符号能与哪些字母结合①。此外，常用词和其他标志也可以成为语种识别的依据。

首先谈谈附加字母的问题。我们都知道，拉丁字母共有26个基本字形（a, b, c, d, …, x, y, z)，由于各种语言都有各自的语音上的特点及不同的拼写规则，于是就数量不等地采用了附加字母。事实上，在使用拉丁字母的各种文字中，完全不采用附加字母的语种是极少的。是不是可以说，对于许多语言，26个基本字母不够用呢？基本上可以这样说，但也不尽然。一般来说，一个语言的音值多于 26 个字母，为了表达这些音值，有的语言采用字母组合，有的则采用附加字母，还有的既采用字母组合，又采用附加字母。我们注意到，很多语言在采用附加字母的同时，却把某些基本字母弃而不用，这是很有趣的。缺少某些字母其实也可作为区别特征。例如芬兰语一般只用 19 个字母，缺少 b、c、d、f、g、q、w、x、z，但增加了 ä 和 ö。又如波兰语缺少 q、v、x，但又引入了 ą、ć、ę、ł、ń、ó、ś、ż、ź 等附加字母。

构成附加字母的最常用的方法就是在基本字母的上面、中间或下面添加一些区别符号。区别符号按形状可分为单点(·)、双点(··)、圈(。)、横(-)、波(~)、勾(ˇ)、折(ˆ)、左撇(ˊ)、右撇(ˋ)、双撇(˝)、弧(⌣)、逗点(,)、反逗点(‘）等等。区

①带区别符号的字母实质上是附加字母。某些语言字母表上虽无这些附加字母，但实际上存在这些字母，因为它们各有自己的音值，书写时也不能够省略，如法语的 élève, 必须写成三个 e: é、è 和 e。在有的打字机上甚至将某些带区别符号的字母独立为一键。以下我们将带区别符号的字母统称为附加字母。

别符号按它们在字母上的位置，还可分为上加符、中加符和下加符。

在使用附加字母的语言中，一个字母一般都只能带一种区别符号。但是越南文的情况比较特殊，时常可以见到一个字母同时带两种区别符号的情况，例如 ậ，ễ，ớ 等等。

附加字母除用区别符号构成之外，还包括一些新造字母。有的新造字母是两个基本字母的合成形式，例如 æ；有的是基本字母的变体形式，例如 þ。

下面让我们用一张表来总结一下在给定的语种范围内的各种区别符号和附加字母（参见表 1）。

给定语种为：阿尔巴尼亚文（Al）、加泰隆文（Ca）、捷克文（Cz）、丹麦文（Da）、荷兰文（Du）、英文（En）、世界语（Es）①、芬兰文（Fi）、法文（Fr）、匈牙利文（Hu）、德文（Ge）、冰岛文（Ic）、印度尼西亚文（In）、意大利文（It）、拉丁文（La）、卢萨文（索布语）（Lu）②、挪威文（No）、波兰文（Pl）、葡萄牙文（Pt）、罗马尼亚文（Ru）、塞尔维亚文（Se）③、斯洛伐克文（Sk）、斯洛文尼亚文（Sn）、西班牙文（Sp）、瑞典文（Sw）、土耳其文（Tu）、越南文（Vi）。

在上述 27 种语言中，英文和印尼文完全不用区别符号；拉丁文有一套元音字母的标音符号，但有时用，有时不用；荷兰文在元音字母上面偶或添加〔ˊ〕、〔ˋ〕、〔¨〕等区别符号，但由于用得很少，难以形成固定特征。除这四种语言之外，其他 23 种语言都是有规律地采用某些区别符号，构成各自的附加字母。事实上，一种语言可同时使用几种附加字母，而一种附加字母又可

①世界语（Esperanto）是波兰眼科医生柴门霍夫博士（1859—1917）在1887年公布的一种人造的国际辅助语。

②卢萨语（Lusatian）是流行于东德的最东南部——卢萨西亚的一种斯拉夫语，又称索布语。

③全称塞尔维亚-克罗地亚文，同时采用两种字母表。一种是拉丁字母表，另一种是斯拉夫字母表。这里介绍的是使用拉丁字母的塞尔维亚文。

表 I　区别符号和附加字母一览表

附加字母 / 区别符号 \ 基本字母		a	e	i	o	u	c	d	g	h	j	l	n	r	s	t	y	z
上加符	˙																	ż
	¨	ä	ë	ï	ö	ü												
	˚	å				ů												
	¯	ā	ē	ī	ō	ū												
	~	ã			õ								ñ					
	ˇ		ě				č	ď					ň	ř	š	ť		ž
	^	â	ê	î	ô	û	ĉ		ĝ	ĥ	ĵ				ŝ			
	´	á	é	í	ó	ú	ć						ń	ŕ	ś		ý	ź
	`	à	è	ì	ò	ù												
	˝				ő	ű												
	˘	ă	ĕ	ĭ	ŏ	ŭ			ğ									
下加符	¸						ç								ş	ţ		
	˛	ą	ę															
中加符	-							đ										
	/				ø							ł						
	’							d’				l’				t’		
新造字母	合成性	æ			œ													
	变体性			ı				ð							β	ꝑ		

同时被几种语言所共用。下面我们用一张矩阵表来标示这两种关系。凡是当某种语言中包括某种附加字母时，一般在其交叉处标一个“+”号。如果某一个附加字母仅用于某一种语言的话，则用“#”号标示。如果某一个附加字母在某一语言中不稳定地隐现，则在标记外面加上括号。由于英文和印尼文不使用附加字母，所以矩阵表中不包括这两种语言。荷兰文只以字母 e 作局部特征记载。越南文由于以特殊的方式使用区别符号，已经构成区别特征的也不予标示(参见表Ⅰ)。

当一种附加字母仅为一种语言所用时，它当然可以成为识别该语种的标志，例如德文的 β ，西班牙文的 ñ，捷克文的 ů，世界语的 ŝ 等等(参见表Ⅰ中带#号的标本)。当一种附加字母被几种语言共同采用时，虽然缩小了范围，但仍不能据此确定语种。在这种情况下，附加字母的组合方式就成为语种识别的重要依据。所谓附加字母的组合方式，是指某种语言所用附加字母的范围以及这些附加字母在行文中所体现的规律。在矩阵表中我们可以看到，附加字母ä为芬兰文、德文、斯洛伐克文和瑞典文所共用。当我们在某一文本中发现有字母ä出现时，我们首先可以把范围缩小到这四个语种。然而究竟是哪种语言呢？这就需要我们比较这四种语言的附加字母的组合方式了。通过查矩阵表我们还可知道这四种语言的附加字母使用范围如下：芬兰文：ä、ö；德文：ä、ö、ü、β；瑞典文：ä、ö、å；斯洛伐克文：ä、á、č、d'、(ď)、é、í、l'、ň、ó、ô、ŕ、š、t'、(ť)、ú、ý、ž。

我们看到，以上四种语言虽然都使用ä，但它们的附加字母域各不相同。斯洛伐克文虽然使用 16 种附加字母，但却不包括 ö，这样就可以把斯洛伐克文同其他三个语种区分开来。剩下的三个语种都同时具备 ä 和 ö，如果尚发现有 ü 或 β 则无疑是德文，如果还有 å 则肯定是瑞典文，如果没有发现 ä 与 ö之外的其他附加字母，那么在我们给定的语种范围内则只能是芬兰文了。事实上，芬兰文还有独特的附加字母的排列方式，我们

表 I　附加字母应用情况矩阵表

附加字母 \ 语种	Al	Ca	Cz	Da	Du	Es	Fi	Fr	Ge	Hu	Ic	It	La	Lu	No	Pl	Pt	Ru	Se	Sk	Sn	Sp	Sw	Tu	Vi
ä							+		+											+			+		
å				+											+								+		
ā													(#)												
ã																	+								+
â								+									+							+	+
á			+							+	+						+			+		+			+
à		+						+				+					+								+
ǎ													(+)					+							+
ą																#									
æ				+							+		(+)		+										
ë					(+)			+									+								
ē													(#)												
ě			+											+											

表 I （续）

附加字母 \ 语种	Al	Ca	Cz	Da	Du	Es	Fi	Fr	Ge	Hu	Ic	It	La	Lu	No	Pl	Pt	Ru	Se	Sk	Sn	Sp	Sw	Tu	Vi
ê								+									+								+
é		+	+		(+)			+		+	+	+					+			+		+			+
è		+			(+)			+				+					+								+
ě													(#)												
ę																#									
ï		+					-	+									+								
ī													(#)												
î								+				+						+						+	
í		+	+							+	+	+					+			+		+			+
ì												+													+
ĭ													(#)												
ı																								+	
ö							+		+	+	+												+	+	

表 I （续）

附加字母 \ 语种	Al	Ca	Cz	Da	Du	F_3	Fi	Fr	Ge	Hu	Ic	It	La	Lu	No	Pl	Pt	Ru	Se	Sk	Sn	Sp	Sw	Tu	
ō													(#)												
õ																	+								+
ô								+									+			+					+
ó		+	+							+	+	+		+		+	+			+		+			+
ò		+										+					+								+
ő										#															
ŏ													(#)												
œ								+					(+)												
ø				+											+										
ü		+						+	+	+							+					+		+	
ů			#																						
ū													(#)												
û		+	+					+																+	

表 I （续）

附加字母＼语种	Al	Ca	Cz	Da	Du	Es	Fi	Fr	Ge	Hu	Ic	It	La	Lu	No	Pl	Pt	Ru	Se	Sk	Sn	Sp	Sw	Tu	Vi
ú										+	+	+					+			+		+			+
ù								+				+													+
ű										#															+
ŭ						+							(+)												
č			+											+					+	+	+				
ĉ						#																			
ć														+		+			+						
ç	+	+						+									+							+	
ď			+																	+					
d'			+																	+					
đ																			+						+
ð											#														
ĝ						#																			

表 I （续）

语种 / 附加字母	Al	Ca	Cz	Da	Du	Es	Fi	Fr	Ge	Hu	Ic	It	La	Lu	No	Pl	Pt	Ru	Se	Sk	Sn	Sp	Sw	Tu	Vi
ğ																								#	
ĥ						#																			
ĵ						#																			
ł														+		+									
l'																				+					
ñ																						#			
ň			+																	+					
ń														+		+									
ř			+											+											
ŕ																				#					
š			+																+	+					
ŝ						#																			
ś																#									

表 I （续）

附加字母 \ 语种	Al	Ca	Cz	Da	Du	Es	Fi	Fr	Ge	Hu	Ic	It	La	Lu	No	pl	pt	Ru	Se	Sk	Sn	Sp	Sw	Tu	Vi
ş																		+						+	
β									#																
ť			+																	+					
t'			+																	+					
ţ																		#							
þ											#														
ý			+								+									+					+
ż																#									
ž			+											+					+	+	+				
ź														+											

可以较经常地发现 ää、öö、äö 等字母组合，这类特点也可以成为语种识别的一种依据。

除利用附加字母识别语种之外，常用词和其他特征也可以借以识别语言。例如可用 the、and、to 等常用词把英语同荷兰语、印尼语等区分开来，可用 à、au、les、une 等常用词把法语同西班牙、葡萄牙语等区分开来。其他特征的例子有：德语的名词一律大写，印尼语的名词复数用双写或平方符号 [^{2}] 表示，加泰隆文在两个 l 中间有时加点(l·l)，西班牙语有句首的倒问号（¿）和倒叹号（¡）等等。

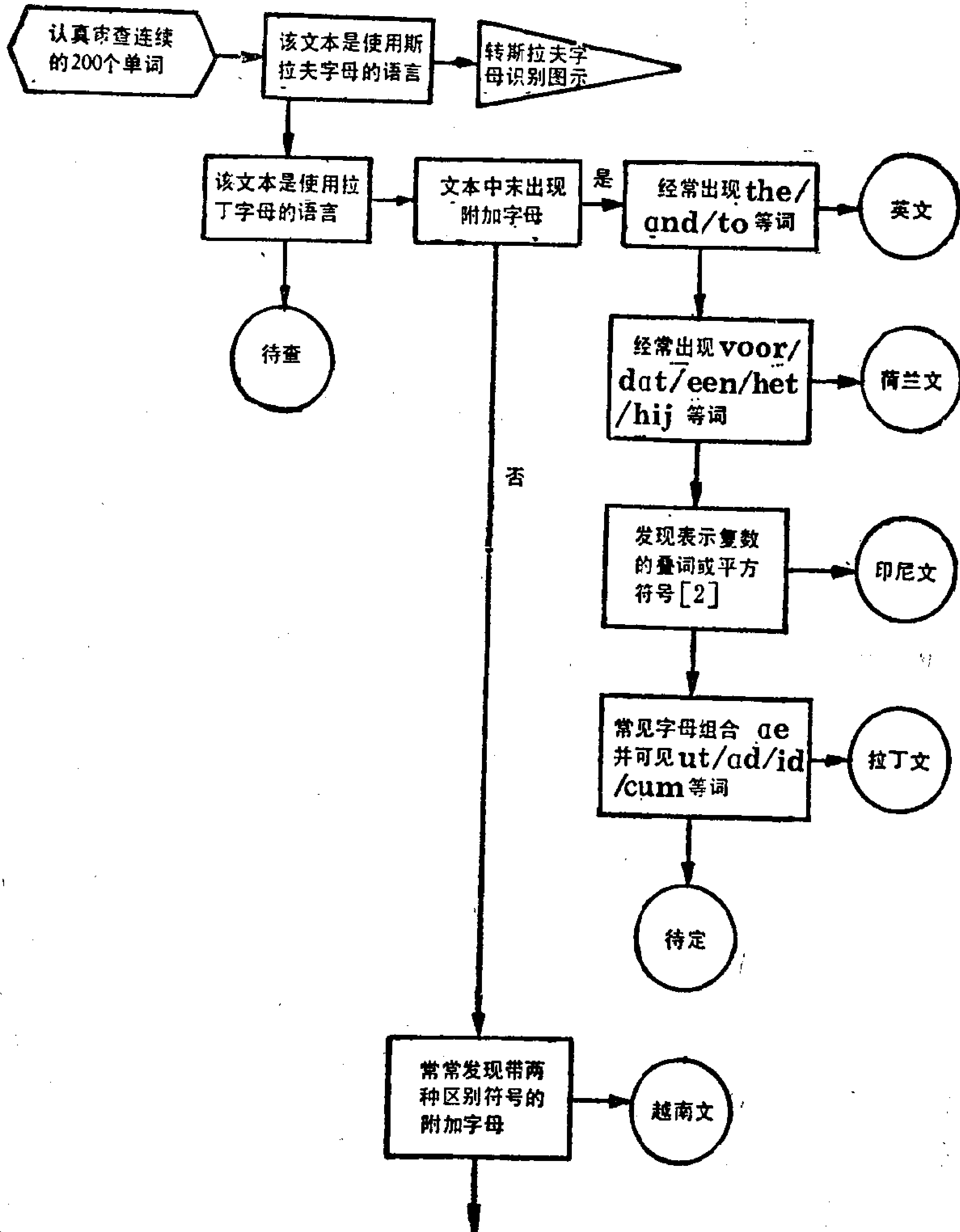
通过上述分析，我们利用语种识别的各种技巧建立了一种程序算法，试图使语种识别更加系统化、规则化和精确化。这种算法用逻辑框图表示(参见图 I)。下面谈谈这种框图的使用方法。

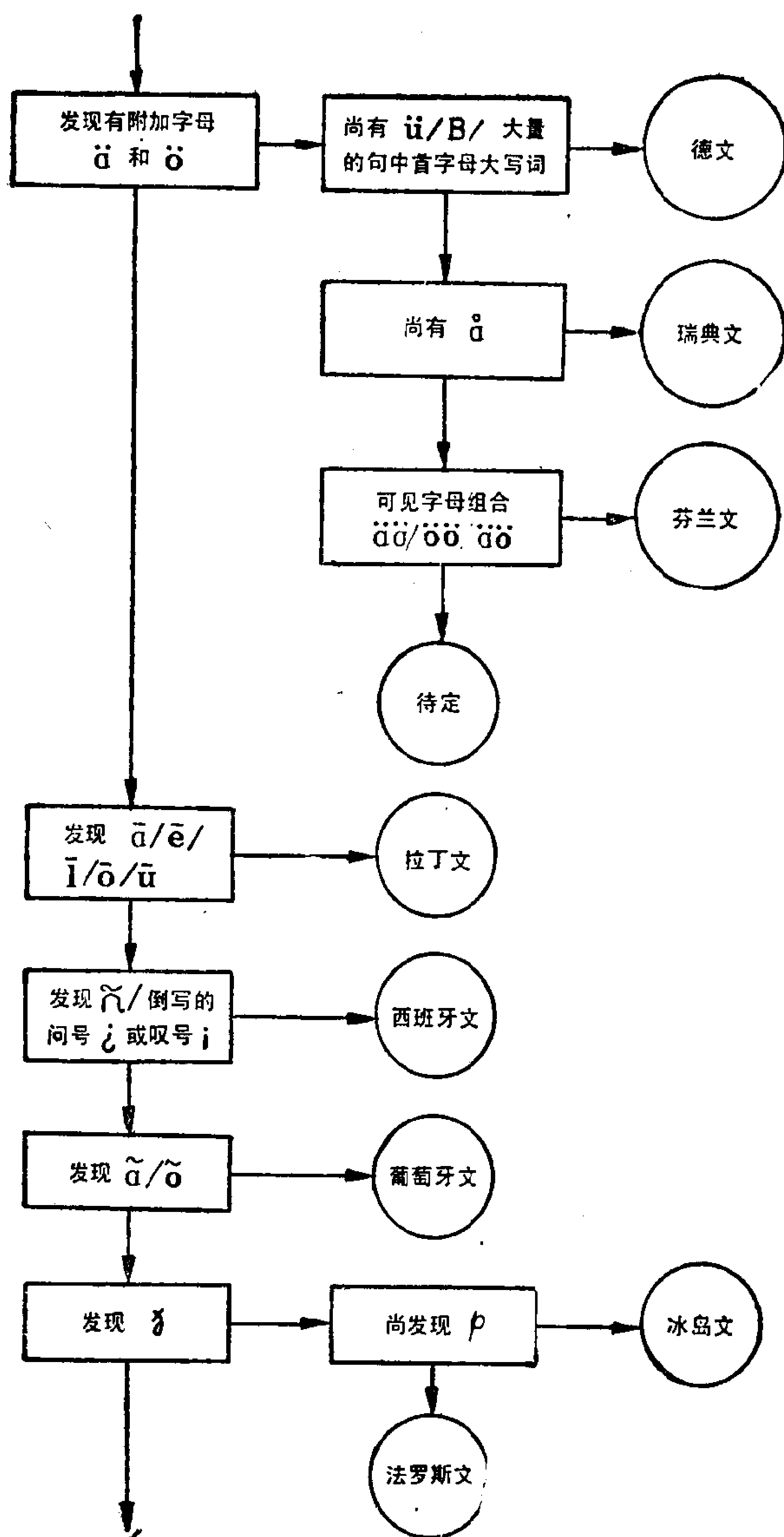
如果有一种使用拉丁字母的文本，若想判断一下它是哪一个语种，那么首先要依次考察足够数量的单词，把附加字母域和其他特征都记录下来。为了确保观察到某语种的各种特征，所考察的原文单词不可少于 100 个，最好能考察 200 个。使用这种框图时，要按从上到下，从左到右的方向进行分析。如果某方框里所陈述的情况与考察的事实相符，则向右走一框，即按“是”的指向走；如不相符，则向下走一框，即按“否”的指向走。如果分析线路已经走到框图里，就说明已经完成了一种语种的鉴定。在给定的语种范围内，只要按照规定进行图示分析，一般是不会走到“待定”框中去的。

此外，还有一些具体问题。例如一个逻辑方框中的陈述是“发现有 ä 和 ö”，这就是说不仅发现了 ä，而且还发现了 ö 时才可右行；如果只发现了 ä 或只发现了 ö 的话，则只能下行。“和”、“及”、“并”、“与”等字都要求几种条件的同时具备。另外，我们用斜杠“/”表示“或者”，例如说发现 ţ/ă，就是说无论发现了 ţ 还是发现了 ă 都可右行，只有这两个附加字母哪一个都没发现时才需下行。

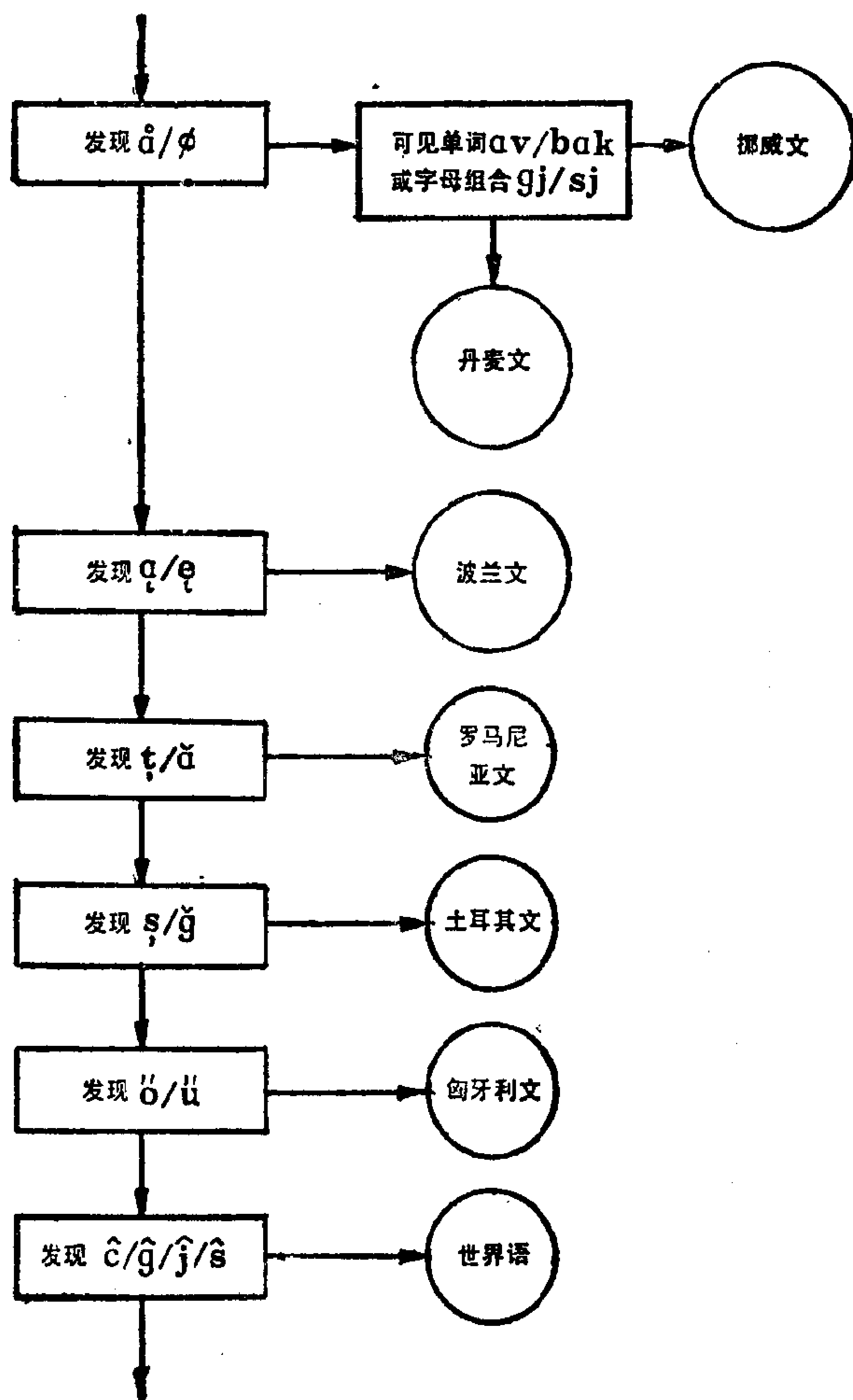
图 I 语种识别逻辑框图

（右行→为是转，下行↓为否转）

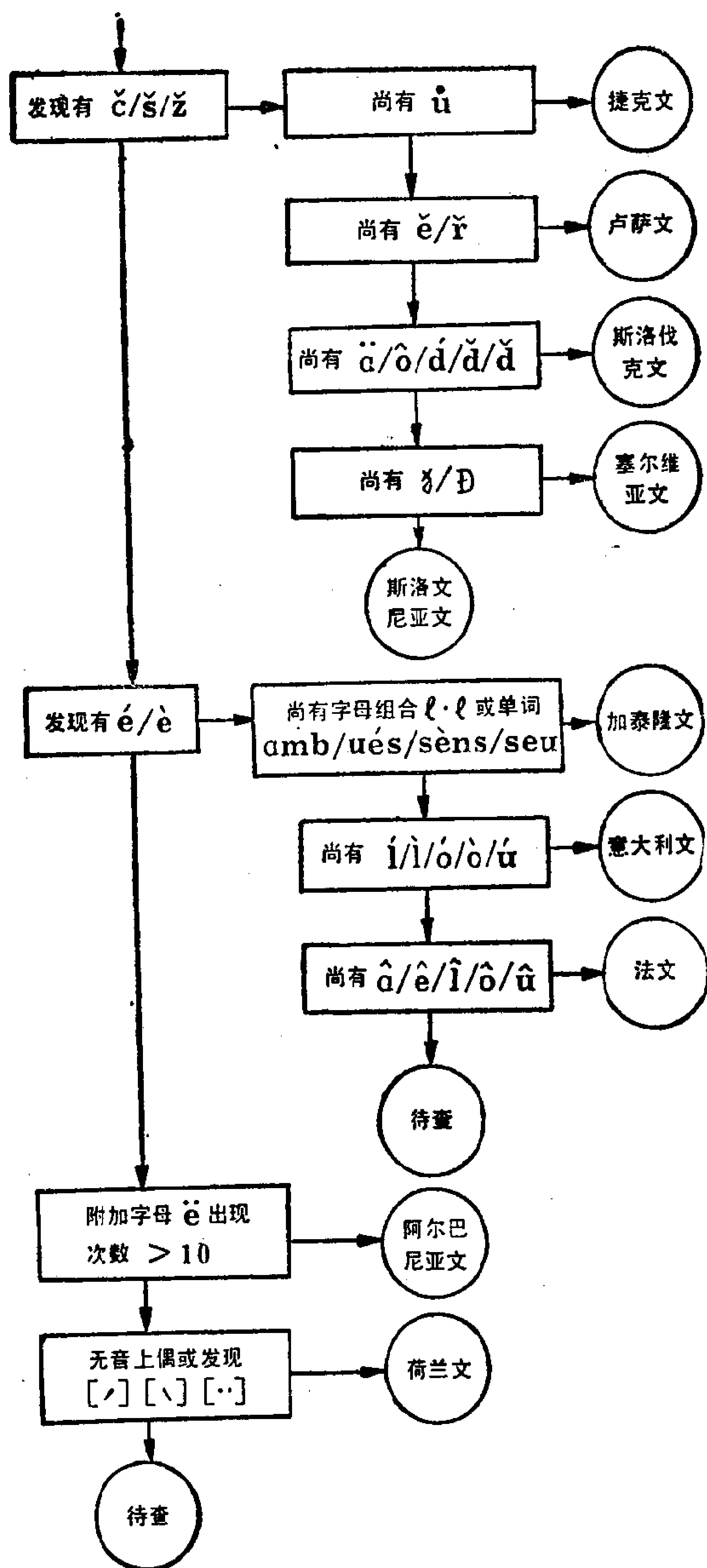




发现有附加字母
ä 和 ö
尚有 ü/ß/ 大量
的句中首字母大写词
德文
尚有 å
瑞典文
可见字母组合
ää/öö äö
芬兰文
待定
发现 ā/ē/
ī/ō/ū
拉丁文
发现 ñ/倒写的
问号 ¿ 或叹号 ¡
西班牙文
发现 ã/õ
葡萄牙文
发现 ð
尚发现 þ
冰岛文
法罗斯文



发现 å/ø
可见单词av/bak
或字母组合gj/sj
挪威文
丹麦文
发现 ą/ę
波兰文
发现 ţ/ă
罗马尼亚文
发现 ş/ğ
土耳其文
发现 ő/ű
匈牙利文
发现 ĉ/ĝ/ĵ/ŝ
世界语



发现有 č/š/ž
尚有 ů
捷克文
尚有 ě/ř
卢萨文
尚有 ä/ô/ď/ď/ď
斯洛伐克文
尚有 ǯ/Đ
塞尔维亚文
斯洛文尼亚文
发现有 é/è
尚有字母组合 l·l 或单词 amb/ués/sèns/seu
加泰隆文
尚有 í/ì/ó/ò/ú
意大利文
尚有 â/ê/î/ô/û
法文
待查
附加字母 ë 出现次数 >10
阿尔巴尼亚文
无音上偶或发现 [´] [`] [¨]
荷兰文
待查

由于我们这种语言识别算法是按一定的次序设计的，所以鉴别每一种语言时都要从图示依次分析，直至走到结论框为止。下面给出十种文字样品，请按照图示分析识别一下。

文字样品

样品 1

Mateusz się porwał w ten mig do niego, ale nim mógł zmiarkować co bądź, już Antek skoczył jak ten wilk wściekły, chycił go jedną ręką za orzydle, przydusił aż tamten dech i głos stracił, drugą ujął za pas, wyrwał z miejsca jak kierz, nogą drzwi wywalił na dwór, i poniósł go prędko za tartak, do rzeki ogrodzonej płotem i cisnął z całej mocy, aż cztery żerdki trzasły kiej słomki, a Mateusz niby kloc ciężki padł we wodę.

样品 2

Gdje su dukati, gdje je snaga, da vidimo ? Duza je ustala da pokaže snagu. Krupna udovica, plećata. Glava joj velika, lice široko, blijedo, lijevo oko sasvim zatvoreno S kuka joj na kuk zategla pregača. U'vrh pregače biju pletenice, debcle, kosmate, u njima dinduvi sa ulara, kolo od sata, zub od vuka. U lice joj naide rumen — zastidjela se pred muškarcima. Pride vreći. Bila je vicca puna žita, zašiljčana, zavezana kanafom. Starcu se po bradi prcsu radost: znao je on snagu svoje snahe.

样品 3

Falou então de si, com modéstia: reconhecia, quando via na capital tão ilustres parlamentares, oradores tão sublimes, tão consumados estilistas, reconhecia que era um zero ! — E com a mão erguida formava no ar, pela junção do polegar e do indicador, um O: um *zero* ! proclamou o seu amor à pátria: que amanhã as instituições ou a família real precisassem dele — e o seu corpo,

a sua pena, o seu modesto pecúlio, tudo oferecia de bom grado! Queria derramar todo o seu sangue pelo trono!

样品 4

Um die Araber von der Halbinsel zu verdrängen, hatten vor Jahrhunderten Königtum und Kirche ein unlösliches Bündnis eingehen müssen. Der Sieg war möglich nur, wenn es den Königen und den Priestern gelang, die Völker Spaniens durch strengste Disziplin zusammenzuschweißen. Es war ihnen gelungen. Sie hatten sie vereinigt in einem inbrünstig wilden Glauben an Thron und Altar. Und diese Härte, diese Einheit war geblieben.

样品 5

Katsantokannat ovat sovittamattomat. Mutta tämä sanan vaihto on kiintoisa sikäli, että se antoi taiteilijamestarille aiheen eräiden yleispätevien ajatusten lausumiseen taiteilijan suhteesta kuvattavaansa henkilöön yleensä. Aaltonen kirjoittaa: "Kun Kiven tyyppiä kerta kaikkiaan ei täydellisesti tunneta ja kun kaikki kertomatiedot hänen ulkonäöstään ovat epätäydellisiä ja ristiriitaisia keskenään, ei hänestä, nyt eikä vastaisuudessa voida tehdä näköistä muotokuvaa ..."

样品 6

Đẫy vườn cỏ mọc lau thưa,
Song trăng quanh-quẽ, vách mưa rã-rời.
Trước sau nào thấy bóng người,
Hoa đào năm ngoái còn cười gió đông.
Xập-xè én liệng lầu không.
Cỏ lan mặt, đất, rêu phong dấu giầy.
Cuối tường gai góc mọc đầy.
Đi về nầy những lối này năm xưa.

样品 7

Čím menšie je niečo, tým väčšmi kričí. Taký fafrnok, nevie to ešte ani hovorit', a prekričí celú rodinu, malý úradník urobí viacej štabarcu okolo seba ako desat' ministrov, malý vrabec prečvirikà svojho starého otca pokojne skáčúceho po chodníku, malý vozík bude rapotat' na celú dedinu, kým štyridsat' konských síl prebrnkne vedl'a teba počuješ iba l'ahké ssst.

样品 8

Şi scurt şi cuprinzător, sărut mîna mătuşei, luîndu-mi ziua bună, ca un băiet de treabă; ies din casă cu chip că mă duc la scăldat, mă şupuresc pe unde pot şi, cind colo, mă trezesc în cireşul femeii şi încep a cărăbăni la cireşe în sîn; crude, coapte, cum se găseau. Şi cum eram îngrijit şi mă sileam să fac ce oi face mai degrabă, iaca mătuşa Mărioara c-o jordie in mînă, la tulpina cireşului !

样品 9

Au fond de son âme, cependant, elle attendait un événement. Comme les matelots en détresse, elle promenait sur la solitude de sa vie des yeux désespérés, cherchant au loin quelque voile blanche dans les brumes de l'horizon. Elle ne savait pas quel serait ce hasard, le vent qui le pousserait jusqu'à elle, vers quel rivage il la mènerait, s'il était chaloupe ou vaisseau à trois ponts, chargé d'angoisses ou plein de félicités jusqu'aux sabords. Mais, chaque matin, à son réveil, elle l'espérait pour la journée, et elle écoutait tous les bruits, se levait en sursaut, s'étonnait qu'il ne vînt pas; puis, au coucher du soleil, toujours plus triste, désirait être au lendemain.

样品10

Han stod rak — som en snurra sålänge piskan viner. Han var blygsam — i kraft av robusta överlägsenhetskänslor. Han var icke anspråksfull: vad han strävade efter var endast frihet från oro, och andras nederlag gladde honom met än egna segrar. Han räddade livet genom att aldrig våga det — Och klagade över att han icke var förstådd!

识别的结果将会如何？如果认真考察了这十种语言的特征，并一一按框图进行推导的话，恐怕是不会搞错，也不致走到待查的结论中去的。正确的答案是：样品1波兰文，2塞尔维亚文，3葡萄牙文，4德文，5芬兰文，6越南文，7斯洛伐克文，8罗马尼亚文，9法文，10瑞典文。

以上谈了按文本识别语种的问题。也许在另外一些场合，还可能遇到按语种找文本的问题。这个问题也属于语种识别问题，例如想在包括很多语种的书刊中找出某一语种的书。为了解决这一问题，在此归纳总结了各语种突出的区别特征表供您参阅（表Ⅰ）。使用这张表时，建议首先能够了解越南文的特征，及时把它排除。其次，将需要找的语种，认真查得该语种的突出的区别特征，再以这些特征与待查核的各文本一一对照比较，相符合的即是。

现在请参照表Ⅰ，从以下五种文字样品（样品11—15）中挑出荷兰文、匈牙利文和丹麦文。

表Ⅰ 各语种突出的区别特征

语种	突出的区别特征
Al	附加字母 ë 的使用频率极高（在100个单词的文本中，ë 的出现次数不少于10），且经常处在单词末一字母
Ca	双 l 中间有时加点（·），如 l·l/ 使用其他语言所没有的 amb/ués/séns/sen 等单词

表Ⅰ （续）

语种	突出的区别特征
Cz	采用附加字母 ů
Da	很少使用附加字母，在局部文段中未见或偶见在元音字母上有〔´〕、〔`〕、〔··〕等区别符号；使用 voor/dat/een/het/hij 等单词
Du	很少使用附加字母，在局部文段中未见或偶见元音字母上有〔´〕、〔`〕、〔··〕等区别符号；使用 voor/dat/een/het/hij 等单词
En	不使用附加字母，且经常使用 the/and/to/that 等单词
Es	采用附加字母 ĉ/ĝ/ĵ/ŝ
Fi	采用附加字母组合 ää/öö/äö
Fr	采用附加字母 ê 但不用 ã 和 õ/ 使用 a/au/les/une 等单词
Ge	采用附加字母 β/句间首字母大写的单词很多
Hu	采用附加字母 ő/ű
Ic	采用附加字母 ð 和 þ/á/é/í/ó/ú
In	不采用附加字母，且复数用数字符号 [²]（平方号）或词的重叠形式表示。
It	采用附加字母 ì；使用 dai/di/gli/già 等单词
La	有时采用附加字母 ā/ē/ī/ō/ū，有时却完全不同；使用 cum/ex/id/ut/guae 等单词且常用字母组合 æ
Lu	采用附加字母 ė 和 ł
No	采用附加字母 ϕ 且使用 av/Dak/gjennem/mot 等单词
Pl	采用附加字母 ą/ę
Pt	采用附加字母 ã/õ
Ru	采用附加字母 ţ/ 采用附加字母 ş，但不用 ü
Se	采用附加字母 č/š/ž/ 和 đ/Đ。
Sk	采用附加字母 č/š/ž 和 ä/ô

表Ⅱ （续）

语种	突出的区别特征
Sn	除 č、š、ž 之外，不用其他附加字母
Sp	采用附加字母 ñ/ 句首使用倒问号（¿）或倒叹号（¡）
Sw	采用附加字母 ä 和 å
Tu	采用附加字母 ğ
Vi	采用带两种区别符号的附加字母

样品11

Viti më njëzet e pesë a njëzet e gjashtë, s'më kujtohet tamam na erdhën n'Elbasan dy vetura të mëdha. Kujtoni ato veturat e moçme. Kish nj'a dy vjet që ishte, çelë xhadeja e automobilit që lidh Elbasanin me portin e Shqipërisë, Durrsin. Më parë ne udhëtonim vetëm me kalë e më këmbë. Kam qenë nj'a pesëmbëdhjetë vjeç djaië ahere. Punoja si hyzmeqar te një qeraxhi, i cili mbet pa punë, sepse rruga e Durrsit, që u hap për automobilat, i la pa punë shum qeraxhinj.

样品12

For mange år siden levede en kejser, som holdt så uhyre meget af smukke, nye klæder, at han gav alle sine penge ud for ret at blive pyntet. Han brød sig ikke om sine soldater, brød sig ej om komedie eller om at køre i skoven, uden alene for at vise sine nye klæder. Han havde en kjole for hver time på dagen, og ligesom man siger om en konge, han er i rådet, så sagde man altid her: "Kejseren er i klædeskabet!"

样品13

Hun trotse nekken en krullende manen geven hun de allure van het overwinnend driespan in een antieke Romeinse wagenren,

en hun gouden ogen schijnen wijdopengespalkt van verbazing, dat ze met onverklaarbare achterlating van lijf en staart hun roemruchte carrière moesten beëindigen in een straat van een dergelijke sjofelheid ... Toen hebben we honger geleden, maar in de volgende oorlóg hoeft dat gelukkig niet meer, daar zal die bom wel voor zorgen, da 's tenminste één troost. Wat een rotleven tegenwoordig, dat ze 't je zelfs tussen twee ooriogen in niet meer gunnen om behoorlijk te eten en eens even mens te zijn.

样品14

Na de grote Islaenner, de grote schöttenjaeger, sem útlegs: sá stóri islendingur og pilsaveiðari. Einhversstaðar var skelt uppúr. Salgestir lutu hans herradómi um leið og hið háa föruneyti sveif innareftir gólfinu. Islendingurinn hélt áfram að standa einsamall. Og þegar hann leit i kringum sig á aðra gesti lést einginn hafa tekið eftir þvi sem hafði gerst; og enn var honum jafnóijóst og áður hver hann var i augum þessa samkvæmis, eða hvar hann stóð; uns enn uppskýtur við hllð hans þeim feita mjúkmála þýskara frá Hamborg.

样品15

A pokoli komédia még egyre tartott a börzén. A halálra ítélt papírok, a bondavárr gyártelep s a bondavári vasút részvényei egyik kézből a másikba repültek. Most már a komikumig vitték a tragédiát, s kezdett humor vegyülni a szerencsétlenségbe. Ez a szó: "Bondavár" csak arra való volt, hogy derültséget idézzen elő a börzeemberek között. Aki az utolsó részvényén túladhatott nagy veszteséggel, nevetett azon, aki megszererzte azt. Kezdték a részvényeket megfoghatlan becsű tárgyakért cserébe kínálgatni. Ráadásul egy új esernyőre egy ócska esernyőért. Használták előfizetési ekvivalensül olyan lapokért, amiket valakinek a szerkesztő a nyakára köt erővel. Ajándékozgatták jótékony célokra.

答案：荷兰文(样品13)，匈牙利文(样品15)，丹麦文(样品12)。

2.3 斯拉夫字母的语种识别 使用斯拉夫字母的语言虽然没有使用拉丁字母的语言那么多，但仅在苏联境内，就有几十种使用斯拉夫字母的民族语言。除俄语之外，还有白俄罗斯语、乌克兰语、阿塞拜疆语、土库曼语、哈萨克语、乌兹别克语、塔吉克语、吉尔吉斯语、鞑靼语、巴尔基什语、雅库特语、布里雅特语、卡尔梅克语等等。在苏联本土之外，尚有几个国家的民族语言也采用斯拉夫字母，主要有保加利亚语、蒙古语①、塞尔维亚语②、马其顿语。

这里涉及的语种识别问题，我们限定在苏联国内的三种最重要的语言：俄语、白俄罗斯语和乌克兰语，以及苏联境外的上述其他国家的四种语言，共计七种使用斯拉夫字母的语言。一般地说，这个范围已能基本适应我们实际工作的需要了。

首先，我们必须把使用斯拉夫字母的语言同使用拉丁字母的语言区别开来。所谓斯拉夫字母，更确切地说，叫作西里尔字母，是在9世纪由一个名叫西里尔的希腊人创造的，因此得名。如果我们把俄文字母表当作西里尔字母的标准字母表的话，我们看到，在它的33个字母当中，就其印刷体来说，一部分同拉丁字母形同音亦似，一部分同拉丁字母形同音不同，一部分小写字母借用了拉丁字母的大写字母，一部分音、形均同希腊字母，另一部分是独特的西里尔字母(参见表Ⅰ)。显然，除与拉丁字母同形的之外，其他的西里尔字母都具备了与拉丁字母相互辨异的区别特征。借助这些字母，我们可以首先把使用斯拉夫字母的语言同使用拉丁字母的语言区分开来。

---

①这里所说的蒙古语是指蒙古人民共和国的蒙古语，我国境内的蒙古语是采用民族字母书写的。

②这里是指使用斯拉夫字母的塞尔维亚文。

表Ⅰ 西里尔字母的类型

类型＼字体		大写字母	小写字母
与拉丁字母同形的字母	发音大致与拉丁字母相同的	A, K, M, O, T	a, o
	发音差异与拉丁字母较大的	B, E, Ë, H, P, C, X, У	e, ё, c, x, y, (m, g, л, u)①
与希腊字母同形同音的字母		Г, П, Ф	
与大写拉丁字母同形的小写字母			м, т, к, в, н I(仅用于ы)
独特的字母		Д, Л②, Б, З, Ж, И, Й, Ц, Ч, Ш, Щ, Э, Ю, Я.	д, л, б, з, ж, и, й, ц, ч, ш, щ, э, ю, я, ъ, ь, ы

下面讨论我们给定的七种使用斯拉夫字母的语种识别问题。首先探讨这七种文字在形式上的区别特征，如果仍然以俄文字母表为标准进行比较的话，有的语言引入了一些附加字母，有的语言舍弃了一些标准字母。从附加字母的类型来看，有的是借用其他的拉丁字母，有的是在字母上添加区别符号，有的是新造的字母(参见表Ⅱ)。

其次，分析一下这七种语言相对于俄文字母表的字母增舍及使用范围(参见表Ⅲ)。

---

① т, д, л, и 等字母有时印成此状。

② д, л 这两个字母虽然起源于Δ, Λ这两个希腊字母，但由于形状改变较大，不再视为与希腊字母同形。

表Ⅱ　附加字母的类型

拉丁字母型的附加字母	不带区别符号	i, j, s, h
	带区别符号	ќ, ў
西里尔字母型的附加字母	带区别符号	ѓ
新造字母	合　成　性	Љ, Њ
	变　体　性	Є, Ө[①], Ә Ү, Џ, Ң, Қ, Ђ/, ђ Ћ/ћ[②]

其次，分析一下这七种语言相对于俄文字母表的字母增舍及使用范围(参见表Ⅲ)。

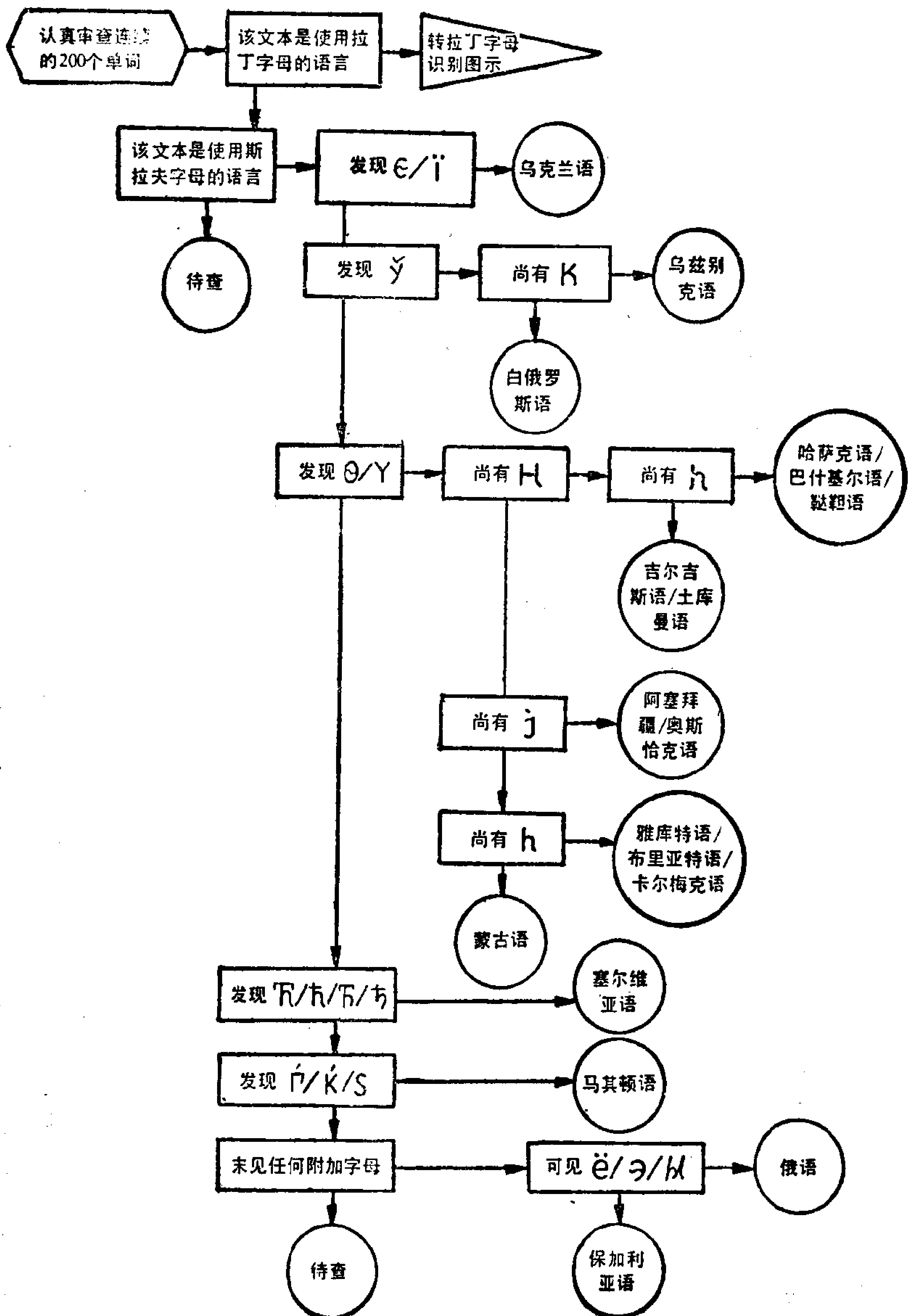
表Ⅲ　七种语言的字母使用范围

字母增舍 / 语种	引入的附加字母	弃而不用的标准字母
乌克兰语	i, ï, є	ё, э, ы
白俄罗斯语	i, ў	—
俄　　语	—	—
蒙古语	Ө, Ү	—
保加利亚语	—	ё, э, ы
塞尔维亚语	j Љ, Њ, Џ, Ђ, ђ, Ћ, ћ	ё, э, ы
马其顿语	j, s, ќ, ѓ, Џ, Љ, Њ	ё, э, ы, й, щ, ь, ъ, ю, я

①字母Ө可转写为拉丁字母的 oh，因此视之为带中加区别符号[～]的附加字母亦无不可。这个字母是元音和谐规则的产物。

② Ђ是ђ的大写，Ћ是ћ的大写。将ћ纳入带区别符号的拉丁字母亦无不可。

图 I　语种识别逻辑框图



认真审查连续的200个单词
该文本是使用拉丁字母的语言
转拉丁字母识别图示
该文本是使用斯拉夫字母的语言
发现 Є/Ї
乌克兰语
待查
发现 ў
尚有 Қ
乌兹别克语
白俄罗斯语
发现 Ө/Ү
尚有 Ң
尚有 ң
哈萨克语/巴什基尔语/鞑靼语
吉尔吉斯语/土库曼语
尚有 ј
阿塞拜疆/奥斯恰克语
尚有 h
雅库特语/布里亚特语/卡尔梅克语
蒙古语
发现 Ђ/ћ/Ђ/ђ
塞尔维亚语
发现 Ѓ/Ќ/Ѕ
马其顿语
末见任何附加字母
可见 ё/э/ы
俄语
保加利亚语
待查

表Ⅱ已刊出了这七种语言相互之间的区别特征，但考虑到被识别的语种可能会超出给定的范围，为了准确无误地识别这七种语言，我们必须确保区别特征在更大范围中的运用性。最突出的例子是蒙古语的区别特征问题，从表Ⅱ可见，θ 和 γ 这两个附加字母可以把蒙文从这七种语言中区分开来，但是稍稍超出这七种语言的范围，就会发现还有其他十来种语言也都采用 θ 和 γ 这两个附加字母。因此，决不能见到某文中有 θ 和 γ 就武断地认为该文本是蒙古文。事实上，在使用 θ 和 γ 的十来种语言中，只有蒙古语不再采用除这两个字母之外的其他附加字母。考虑到 θ、γ 这两个字母在阿尔泰语系的语言中使用较广，在此也就顺便识别一下与 θ、γ 有关的问题。不过，对于给定范围之外的语言，我们不准备加以仔细识别，因而也不一定给出唯一确定的结论。一般说来，在经常工作中除这七种语言之外，遇到其他使用斯拉夫字母的语言的机会毕竟是很少的。下面给出这七种语言的语种识别逻辑框图(参见图Ⅰ)，其标记规则和加工方法与使用拉丁字母的语种识别框图相同。

由于使用斯拉夫字母的语种识别问题比使用拉丁字母的语种识别问题简单得多，所以图Ⅰ的语种识别框图不但可以用于某一文本的语种识别，还可用于在许多语种的文本中找出所需的语种文本。

下面给出十种文字样品，请根据图示分析首先挑出白俄罗斯语，然后再分别指出其他文本的语种。

样品16

Сутринта Огнянов се упьти към града. Той измина балканского гърло и излезе при манастира. На поляната пред манастира, под големите орехи се разхождаше игуменът, гологлав. Той се възхищаваше от утринната хубост на тия романтични места и приемаше на големи глътки свежия живителен въздух на планината. Есен-

ната природа имаше ново обаяние с тия позлатени листове на дърветата, с тия пожълтели кадифени гърбове на Балкана и с тая сладостнонежна повяхналост и меланхолия.

样品17

Име му је било Алија Џевеница: по занимању био је слуга ... Бак је био одличан, играјући се, без по муке, а изгледало је да се одличне оцене у његовој ђачкој књижици ређају саме од себе, по неком утврђеном реду и следу, као конац који се одмотава са клупка ... Ђуте занети у мисли, поцупкују на својим седиштима како аутобус поскаује на неравном друму, те потсећају на путнике задремале у купеу ускотрачног воза који се цео дан клати ужареном Херцеговином, путујући око бескрајног Поповог Поља.

样品18

— Хайр, биз бу дагдагаларни ўртадан кўтаришга харакат килурмиз,—деди Навоий катъий оханг билан.—Гарчи бу мазхабларнинг бирини ўзгасидан афзал кўрмасак хам, улуснинг бирлигини эътиборга олурмиз. Иним, дунёда китоб ўкимокдан, тафаккурдан, шеър айтмокдан ўзга завкбахш машгулот йўкдир. Табиатим кўпрок бу томонга мойил эди. Сокин бир масканда яшаб, бу завк дарёсида сузмокчи эдим. Лекин, менга, маълумингиз, давлатда вазифа бердилар ... Ёлвиз эл ва улус манфаатини назарга олиб, мансабни кабул этдим. Бу муборак юртда килинадиган ишлар бенихоят кўпдир. Бу ишларнинг хар бирига элимиз асрлардан бери ташнадир.

样品19

«Что это ? я падаю ? у меня ноги подкашиваются», подумал он и упал на спину. Он раскрыл глаза, надеясь увидать, чем кончилась борьба французов с артиллеристами, и желая знать, убит или нет

рыжий артиллерист, взяты или спасены пушки. Но он ничего не видал. Над ним не было ничего уже, кроме неба — высокого неба, не ясного, но все-таки неизмеримо высокого, с тихо ползущими по нем серыми облаками.

样品20

Капитализм ба коммунизмын хооронд шилжилтийн мэдээж Үе бий гэдэгт онолын талаар эргэлзээ алга. Энэ Үе нь нийгмийн аж ахуйн энэ хоёр хэв байдлын шинж тэмдэг буюу шинж чанарыг өөртөө хослуулахгүй байж чадахгүй. Энэ шилжилтийн Үе бол сөнөж байгаа капитализм ба сч ҮҮбайгаа коммунизмын хоорондын, буюу Үр Үгээр хэлбэл, ялагдсан боловч устгагдаагҮй байгаа капитализм ба ҮҮссэн боловч бас л огт дорой байгаа коммунизмын хоорондын тэмцлийн е биш байж чадахгҮй.

样品21

Партызаны падхапіліся, як па камандзе, калі маці падышла. Пастарэлая, ссутуленая, у чорнай хустцы — помню, бачу яе і цяпер — Марына падоўгу стаяла каля кожнай труны, узіраючыся ў твары забітых. Не, яна не плакала. Зрэдку варушыліся засмяглыя вусны. Што яна казала? Чытала малітву? Ці слала праклёны забойцам? Я баяўся: ці не памутнеў ад гора яе розум?

样品22

Тече вода в синє море,
Та не витікає:
Шука козак свою долю,
А долі немає.
Пішов козак світ за очі;
Грає синє море,
Грає серце козацькеє,
А думка говорить:

«Куди ти йдеш, не спитавшись ?
На кого покинув
Батька, неньку стареньќую,
Молоду дівчину ?
На чужині не ті люде, —
Тяжко з ними жити !
Ні з ким буде поплакати,
Ні поговорити».

样品23

Ятакхана тып-тын булып кала кӨндӨн кҮп Өлешендә гӨрләп торган ятакханага бындай тынлык килешмәй. Мәктәптә укыусы «художниктәр» тарафынан стенага кумер, карандаш кәләмдәр менән эшләнгән рәсемдәр ҙә ятакхана ӨҫтӨнә монайып карапкалган һымак була.

样品24

За нас, како за народ што успеа да го оформи својот литературен јазик дури во последниве децении, ќе биде многу поучно да знаеме какви биле општо основните карактеристики на развојот на литературните јазици во словенскиот свет. Ќе биде поучно особено поради тоа што ќе најдеме во тој развој ред аналогии со тоа што се случувало кај нас и што ќе можеме да забележиме извесни закономерности таму каде што инаку може да ни се чини дека некоја појава произлегува само од нашата посебна ситуација.

样品25

Меідана јығылан чамаат кезҮнҮ даш булагдан чыхан нова зилләмишди.

Бирдән су сәсләнди, шән бир курулту илә ахмага башлады. Узун илләр су һәсрәтидә олан әһали һәјәчанла бахыр, сусурду. СҮкуту

сујун шаграг сәси позурду. Инсанлар гурумуш сәһрада илкин су кәрмүш јолчулар кими мәфтунлугла бахыр, динмир, санки сујун сәраба чеврилмәчәјиндән горхурдулар. Лакин көрдүкләри һәгигәт иди. Шаирә Натәван догма елинә Шушаја булаг чәкдирмишди. Бирдән һамы чошду, су һәсрәтилә јанан синәләр галхыб енди:

— Јашасын Натәван.

— Вар олсун Хан гызы.

正确答案如下：白俄罗斯文是样品 21；其他文本的语种：样品 16 保加利亚语，样品 17 塞尔维亚语，样品 18 乌兹别克语，样品 19 俄语，样品 20 蒙古语，样品 22 乌克兰语，样品 23 不确定结论——哈萨克语/巴什基尔语/鞑靼语(实为巴什基尔语)，样品 24 马其顿语，样品 25 不确定结论——阿塞拜疆语/奥斯恰克语(实为阿塞拜疆语)。

我们相信，通过这个线路，可以识别出我们所限定的七种使用斯拉夫字母的语言。

语种识别问题是个有用而又有趣的问题。上面总结、归纳出来的规律主要是供人使用的，如果加上语言自动处理的某些手段，将会使识别过程更加严密、有效。微处理机一旦配上光学字符识别装置，这类工作完全可以交给机器来做。

3. 标音、转写和译音

在文献处理中，经常会遇到文字的标音和转写问题。例如有不少专业文献采用拉丁字母标记日文、中文、俄文、希腊文、朝鲜文等等。这样既可以使各语种的文字整齐划一和便于印刷，还能为国际间文字信息的交流提供统一的标准。因此文字的标音和转写问题也是文字自动处理的一个重要课题。

标音（transcription） 是用字母或音标标记语言的方法，分宽式和严式两种。宽式的也叫音位学标音法，可采用本族的拼音字母（或增加一些附加符号）作标记工具；严式的也叫语言学标音法，一般采用国际音标（International Phonetic Alphabet，简称 IPA）作标记工具。标音法可用来标记另一种语言的语音，也可以用来标记方言的语音。在不要求精确表达语音时，采用宽式标音法。但在调查语言和方言时，只能用严式标音法，因为只有这样，才能归纳出一个语言或方言的音位系统。另外，在教人正确发音时，也采用严式标音法，如词典里都用国际音标注音。

转写法（transliteration）是用一种字母表的字符标记另一种字母表的字符。最普遍的转写法是将斯拉夫字母、阿拉伯字母等非拉丁文字系统的文字符号转写成拉丁字母，一般称之为“罗马化”。日语音节字母用其他字母表的字符标记，也是一种转写。转写是不同字母表字符之间的转换，主要目的在于为每个字母或字母组合求出相应的一个字母或字母组合，而不在于求得实际发音，因而，字符转换时常注重形体一致而不注重发音是否相同。例如，граната、нового 和 легкий 中的 г，都以 g 转写。由于

字符之间的转写是唯一的，因而也是可逆的。例如，由 Ленинград 可转写成 Leningrad，但也可还原。

标音和转写的区别可归结为以下三点：

(1)前者是字符对语音，后者是字符对字符。

(2)前者既可以为有文字的语言标音，也可以为无文字的语言标音，同时既可以标记拼音文字，也可标记非拼音文字；但后者只能用于有文字的语言，而且还限于拼音文字之间。

(3)前者是单向的，不可逆的(语音→字符)；后者是双向的，可逆的(字符⇄字符)。

除了标音和转写，还有另一种标记法，这就是译音。译音(transcription by nonlettered characters)是使用非字母文字标记另一种语言的方法。用汉字标记外语的发音，就属于译音。一般文献中只讲标音和转写两种标记法，这不全面。针对汉语和其他使用非字母文字的语言来说，还应有译音这种标记法，理由是：(1)根据一般的定义，非字母文字之间，或非字母文字与字母文字之间不可能有“转写”；(2)非字母文字只有被“标音”这一过程，而无标记字母文字的过程；(3)为了表示用非字母文字标记字母文字这个过程，无论从理论上或从实践上讲，完全有必要设立“译音”这个术语。

下面分别介绍几种常见的标音法和转写法。最后再论述一下译音法。

3.1 用拉丁字母为日语标音 我们知道，日文的书写体系是由汉字和假名组成的。当我们用拉丁字母为日语标音时，不但应掌握假名的读音，还应掌握日语汉字的读音。目前，日本有三套为日语标音的罗马字方案。一为黑本式(ヘボン式)，二为日本式，三为训令式。内容大同小异，但以黑本式为最常用。在下面的表中，我们主要介绍假名的标音，遇到汉字时，可先按其发音注假名再予标音。标音方案如果三式相同，则不重复罗列，如果不同，则首先列出黑本式，其他两式写在括号内。如果其他两式也相同，

则括号内只写一个，如果不同，则在括号内先列出日本式，再列出训令式。

表1　基本假名的标音

あ　a	い　i	う　u	え　e	お　o
か　ka	き　ki	く　ku	け　ke	こ　ko
さ　sa	し　shi (si)	す　su	せ　se	そ　so
た　ta	ち　chi (ti)	つ　tsu (tu)	て　te	と　to
な　na	に　ni	ぬ　nu	ね　ne	の　no
は　ha[1]	ひ　hi	ふ　fu (hu)	へ　he[2]	ほ　ho
ま　ma	み　mi	む　mu	め　me	も　mo
や　ya		ゆ　yu		よ　yo
ら　ra	り　ri	る　ru	れ　re	ろ　ro
わ　wa				を　o[3]
				ん　n
が　ga	ぎ　gi	ぐ　gu	げ　ge	ご　go
ざ　za	じ　ji (zi)	ず　zu	ぜ　ze	ぞ　zo
だ　da	ぢ　ji (di, zi)	づ　zu (du, zu)[4]	で　de	ど　do
ば　ba	び　bi	ぶ　bu	べ　be	ぼ　bo
ぱ　pa	ぴ　pi	ぷ　pu	ぺ　pe	ぽ　po

二合拗音假名由一个い段假名及其后小写的复元音ゃ、ゅ、ょ拼成。其标音方法，一般由各い段假名罗马字的辅音部分分别加 ya, yu, yo 表示，但尚有一些例外，详见下表：

①は用作格助词时，标音为 wa。
②へ用作格助词时，标音为 e。
③を仅用于格助词，标音为 o。
④ぢ和じ的标音相同，づ和ず的标音相同(另见は—わ，あ—お，へ—え等)，这时，原语与标音之间并非严格的一一对应。

表 2　二拗合音假名的标音

きゃ　kya	きゅ　kyu	きょ　kyo
しゃ　sha (sya)	しゅ　shu (syu, shu)	しょ　sho (syo, sho)
ちゃ　cha (tya)	ちゅ　chu (tyu)	ちょ　cho (tyo)
にゃ　nya	にゅ　myu	にょ　nyo
ひゃ　hya	ひゅ　hyu	ひょ　hyo
みゃ　mya	みゅ　myu	みょ　myo
りゃ　rya	りゅ　ryu	りょ　ryo
ぎゃ　gya	ぎゅ　gyu	ぎょ　gyo
じゃ　ja (zya)	じゅ　ju (zyu)	じょ　jo (zyo)
ぢゃ　ja (zya)	ぢゅ　ju (zyu)	ぢょ　jo (zyo)
びゃ　bya	びゅ　byu	びょ　byo
ぴゃ　pya	ぴゅ　pyu	ぴょ　pyo

无论是基本假名还是二合拗音假名都可构成长音。在日语原文中，长音一般借助于重复あ行各元音组成。お段各行基本假名的长音既可重复お，也可重复う，但标音都一样，拗长音由二合拗音再加上あ或う（不用お）组成。长音罗马字的标音方法很简单，只须在元音上戴个“帽子”既可，以上两表中的各音节均可采用此法构成其长音，具体地说，就是在元音字母上加长音符“ ˆ ”或“－”，从而构成 â, î, û, ê, ô，或者。ā, ī, ū, ē, ō, 除戴帽法之外，长音也可以用双写表示。如 aa, ii, uu, ee(或 ei), oo。值得注意的是，有时两个元音之间有词素界限，例如おもう(omou)中的もう，ながあめ (nagaame) 中的があ，则不能视为长音，也不能使用长音符号为其标音。下面是长音和拗长音的标音例子：

ああ â, たあ tâ, きい kí, しい shî(sî), つう tsû(tû), そお sô,

そう sô, ふう fû(hû), やあ yâ, ちい jî(dî, zî), ぽお pô, ぽう pô。

きゃあ kyâ, しゃあ shâ(shâ, syâ), しょう shô (syô, shô), ちゆう chû(tyû), りょう ryô, じゃあ jâ(zyâ), じゆう jû (zyû), ぢう jô(zyô), びゆう byû。

在日文中，促音发生在さ、た、か、ぱ诸行辅音之前，用小一号的つ表示。为促音标音时，双写该行的辅音即可，例如た行前的促音一般拼作 tt(ち前促音亦可拼作 tch)。

摄氏(せつし) sessi,　1顿(いつとん) itton,　血沈(けつちん) ketchin,

学校(がつこう) gakkô,　日本(につぽん) nippon。

关于拨音ん的标音有一点需要注意，即是当拨音在复合元音や、ゆ、よ之前时，为了起隔音作用，以 n' 标音。例如きんゆう[金融]的标音是 kin'yû。

日文中采用片假名拼写外来语时，常会出现比较特殊的假名组合。对此，在表 3 中列出它们的标音规则。

表 3　特殊假名组合的标音

イェ	ye	ウァ	wa	ウィ	wi	ウェ	we
ウォ	wo	クァ	kwa	クィ	kwi	クェ	kwe
クォ	kwo	クゥ	kwa	シェ	she (sye)	チェ	che (tye)
ツァ	tsa	ツイ	tsi	ツェ	tse	ツォ	tso
ティ	ti (t'i)	テユ	tyu (t'yu)	トウ	tu (t'u)	ニェ	nye
ヒェ	hye	ファ	fa	フィ	fi	フェ	fe
フォ	fo	フェ	fyu	ミェ	mye	ヴ	vu
ヴァ	va	ヴィ	vi	ヴェ	ve	ヴォ	vo
グァ	gwa	グィ	gwi	グェ	gwe	グォ	gwo
ジェ	je (zye)	ディ	di (d'i)	デユ	dyu (d'yu)	ドウ	du (d'u)
ビェ	bye	ピェ	pye	ヴュ	vyu		

在片假名中，用“ー”表示其前元音为长音。为片假名标音时，则需将“ー”转换成为拉丁字母上的长音符号，例如：

ビール bîru, テーブル têburu, トースト tôsuto

下面给出标音实例概略说明上述标音规则。例如：

1. 鈴木(すずき)さんは中国語(ちゅうごくご)があまり上手(じょうず)ではありません。
Suzuki san wa chûgokugo ga amari jôzu dewa arimasen

2. この小学校(しょうがっこう)は昔(むかし)はもっと小(ちい)さかったです。
kono shôgakkô wa mukashi wa motto chiisakatta desu

3. 昔は駅(えき)の附近(ふきん)も静(しず)かだったのですが今(いま)はすっかりにぎやかになりました。

Mukashi wa eki no fukin mo shizuka datta no desu ga ima sukkari nigiyaka ni narimashita。

4. レーダ rêdâ 或 reidâ

テレビジョン	terebijon
チャージ	châji
ヴォルトメーター	vorutomêtâ
オーストラリア	ôsutoraria
カンボジア	kanbojia
イクオール	ikwôru
スティーム	sutîmu
セロファン	serofan
フェット	fetto
フューズ	fyûzu
レヴュー	revyû

3.2 汉语的标音问题 汉语的书写形式是汉字，它是一种表意文字。为了给汉字标音，国内外制订的标音方案很多。其中，大家十分熟悉的且具有权威性的标音体系，是我国1958年颁布的汉语

拼音方案(简称 PY)。汉语拼音方案是一种最佳的标音转写方案，它不仅是我国的法定方案，而且也得到了国际公认。1977年9月，联合国第三届地名标准化会议通过决议，采用它作为中国地名罗马字母拼写法的国际标准；1982年8月，国际标准化组织(ISO)发出ISO—7098国际标准文件，规定它作为世界文献工作中拼写有关中国的专门名称和语词的国际标准。有关汉语拼音方案的材料很多，许多字典和词典也把它列为附录，这里就不作具体陈述了。

既然汉语拼音方案是法定方案和唯一的标准形式，那末是否还有必要介绍别的方案呢？我们说有必要。理由是：(1)汉语拼音方案是一种标准形式，这是针对现在和未来而言的。在此之前，曾经用过很多种标音方案，其中有些在过去的出版物中广泛应用过。(2)由于种种原因，一些国家或地区，或其中的某些部门一时还不能采用这种标准，需要有一个过渡阶段，因而在此期间我们还会遇到其他标音法。其他标音方案还有不少，下面仅介绍四种较为重要的。它们是：我国赵元任等创制的国语罗马字(简称国罗)，德国 F. Lessing 与 W. Othmer 共同创制的方案(简称德式)，英国 Tomas Francis Wade 研制的威妥玛式拼音方案(或称英式)，以及法国 Vissière 创制的拼音方案(简称法式)。

3.2.1 国语罗马字 国语罗马字拼音方案是在钱玄同的倡议下由赵元任等人创制的，1928年经大学院公布，其特点是用字母的变化表示不同的声调。下面分别介绍国罗的基本式和变调韵母。

国罗的基本式包括声母和阴平韵母表。国罗的声母有不少同PY的字形与读音完全一致，它们包括 b, d, f, g, h, k, l, m, n, p, r, s, t 共13个。在国罗中，还有一些声母在表示某一特定音素时与 PY 采取不同的字形，它们包括(括号外是国罗，括号内是 PY)：ts(c), tz(z), i(y), u(w), lh(l), mh(m), nh(n), rh(r)。在国罗中 lh, mh, nh, rh 仅用于阴平音节，其他声调

用 l, m, n, r 拼写①。国罗还有三个兼职声母，它们是 j(zh, j), ch(ch, q), sh(sh, x)。由于 j, ch, sh, 在表示 (j, q, x) 等音时，后面总跟着个 -i，而在表示 (zh, ch, sh) 时，却从不跟着 -i，所以可以兼职。

国罗的基本韵母(阴平韵母)与 PY 一致的有：a, ai, an, ang, e, ei, en, eng, ia, ian, iang, ie, in, ing, iong, o, ong, ou, u, ua, uai, uan, uang, uo。国罗与 PY 字形或用法有区别的韵母有：au(ao), el(er), y(zhi, si … 的韵母 i), i(mi, qi… 的韵母 i，不兼作 zhi, si… 的韵母), iau(iao), iou(iu), iu(ü 或 xu, qu…的韵母 u), iuan(üan 或 quan, xuan…的韵母 uan), iue (üe 或 xue, que 的韵母 ue), uei (ui), uen (un), iun (ün 或 jun, yun… 的韵母 un), è(ê)。

国语罗马字采用复杂的字母变化来表示不同的声调。变调规则详见表 4。我们看到，声调变化的一般规则是：阴平用基本式，阳平韵母加 r，上声双写主要母音，去声音节尾加 h；然而，不合一般规则的例外仍然是大量的。以表 4 的变调规则为基础，还有一种后来发展形成的简化变调规则，即变调一律采用上述的一般规则，毫无例外。林语堂在他所编的汉英词典中即是用这种简化的国罗为汉语标音的。

表 4　国语罗马字变调字母表

阴　平	阳　平	上　声	去　声
a	ar	aa	ah
ai	air	ae	ay
an	arn	aan	ann
ang	arng	aang	ang

①以 l, m, n, r 为声母的音节，阴平与阳平的区别不靠韵母变调，而靠声母变调。阴平声母用 lh, mh, nh, rh, 阳平声母用 l, m, n, r, 韵母用基本式，例如：mha (妈), ma (麻)。

表 4 （续）

阴 平	阳 平	上 声	去 声
au	aur	ao	aw
e	er	ee	eh
ei	eir	eei	ey
el	erl	eel	ell
en	ern	een	enn
eng	erng	eeng	eng
i	yi	ii	ih
ia	ya	ea	iah
ian	yan	ean	iann
iang	yang	eang	iang
iau	yau	eau	iaw
ie	ye	iee	ieh
in	yn	iin	inn
ing	yng	iing	ing
iong	yong	eong	iong
iou	you	eou	iow
iu	yu	eu	iuh
iuan	yuan	euan	iuann
iue	yue	eue	iueh
iun	yun	eun	iunn
o	or	oo	oh
ong	orng	oong	ong
ou	our	oou	ow
u	wu	uu	uh
ua	wa	oa	uah
uai	wai	oai	uay
uan	wan	oan	uann
uang	wang	oang	uang
uei	wei	oei	uey
uen	wen	oen	uenn
ueng	weng	oeng	ueng
uo	wo	uoo	uoh
y	yr	yy	yh

下面是国语罗马字的标音实例。

Jonggwo shyh shyhjieh shang renkoou tzuey duo de gwojia, chahbuduo jann chyuan-shyhjieh renkoou de syh-fen-jy-i。

（中国是世界上人口最多的国家，差不多占全世界人口的四分之一。）

3.2.2 德式汉语标音方案 德语国家的汉学家为汉语标音创制过不少方案，其中流行最广的就是 F. Lessing 和 W. Othmer 共同研制的方案。这个方案最初见于两个创制人合编的《汉语北方话口语教材》之中，用来为汉语标音。此后，此方案广泛用于德语书刊、词典和其他工具书中。在我国出版的德语出版物中也曾使用过此方案。

此方案最显著的特征在于保留了大量的德文拼音规则。下面将其声、韵母表与汉语拼音（PY）的异同分述如下。

先谈谈声母表。德式方案的声母有一些同 PY 的字形与读音完全一致，它们包括：b, d, f, g, h, k, l, m, n, p, t, w, y。另外一些德式声母在表示某一特定音素时采用与 PY 不同的字形，它们包括（括号外的是德式拼音，括号内的是 PY）：dj(j), ds(z), dsch(zh), hs(x), j(r), sch(sh), tj (q), ts(c), tsch(ch), ss(s)。

德式标音方案的韵母与 PY 韵母相一致的有：a, ai, an, ang, ia, iang, in, ing, iu, ou, u, ua, uai, uan, uang, ui, un。与 PY 字形或用法不一致的有：än(ian 中的 an) au(ao), ö (she 的韵母 e), ä(ê 或 xie 的韵母 e), e(ei), ĕn(en), ĕng (eng), örl(er)、ï(zhi, si 的韵母 i), i(mi 的韵母 i, 不兼用), iän(ian), iau (iao), iä(ie), iung(iong), o(o, uo, e), ung (ong), ü(ü 或 xue 内省略两点的 u), üan(quan 内省略 ü 上两点的 un)。

对下面的几个不同点需要特别加以说明：（1）德式方案的韵

母 i 可单独使用，与 PY 的 yi 等价；(2)德式方案的 yu 相当于 PY 的 you；(3)德式方案的 o 兼 PY 的三个音质：o, uo, e；(4) ui 和 ue 都等价于 PY 的 ui，但 ue 仅与 g, k 相拼；(5) ö 和 o 都等价于 PY 的 e，但 o 与 g, k, h, n, l, j (这个 j 等于 PY 的 r)相拼，ö 与其他声母相拼。

下面是用德式标音方案为一些汉语词汇标音的例子(德式方案没有单独的标调系统，在必要时，可用数字标调)：

1. 红旗 hungtji； 2. 政治 dschĕngdschĭ； 3. 合作社 hodsoschö； 4. 后天 houtiän； 5. 旅馆 lüguan； 6. 侵略 tjinlüä； 7. 世界 schĭdjiä； 8. 热水瓶 joschuiping； 9. 降落 djianglo； 10. 邮票 yupiau。

3.2.3 威妥玛式汉语标音方案　在英语国家中，为汉语标音制订的方案也有相当数量，但真正延存下来的只有威妥玛方案。这个方案最初是由 Thomas Francis Wade 创制的，并用于他与 W. C. Hillier 合编的《Yü Yên Tzŭ Êrh Chi (语言自迩集)》(1867)中，用来为汉字标音。后来，剑桥大学汉语教授 Herbert Allen Giles 对方案进行了某些简化，例如去掉零声母 ng，归并 hui 与 huei、hsüan 与 hsüen 等音节，这一改进方案可见于他1892年主编的《汉英词典 (A Chinese-English Dictionary)》；后来，这个方案经过进一步的简化和修正。在 R. H. Mathews 所编的汉英词典(1931)中，威妥玛标音方案最后定型。这种定型的威妥玛方案至今还在各类外语出版物中应用。在西方，每个汉学家都必须掌握这种方案。我国邮政界曾长期使用略加变动的威妥玛式标音方案，只是近年来，由于汉语拼音方案(PY)的普及，威妥玛式方案才逐渐被取代了。

威妥玛方案的下述声母与 PY 一致：f, h, l, m, n, s, sh, w, y。与 PY 不同的声母包括：ch(j, zh), ch'(q, ch), hs(x), j(r), k(g), k'(k), p(b), p'(p), sz(si 的声母 s), t(d), t'(t), ts(z), ts'(c), tz(z), tz'(c)。

威妥玛方案的下述韵母与 PY 一致：a, ai, an, ang, ao, ei, ia, iang, iao, in, ing, iu, ou, u, ua, uai, uan, uang, ui, un。与 PY 不同的韵母有：ê(le 的韵母 e), eh(ê 或 ie 之 e), en (ian 之 an), ên(en), êng(eng), êrh(er), i(mi 之 i 或 yi, 不兼作 si 之 i), ieh(ie), ien(ian), ih(zhi, chi, shi, ri 之 i), iung (iong), o(o, e, uo), yu(you), ŭ(zi、ci、si 的 i), ü(ü 或省略两点的 u), üan (省略 ü 上两点的 uan), üeh(ue), uei(ui), ün(jun 中的 un), ung(ong)。

威妥玛方案有 ê、o 两个字形等价于 PY 的 e, 但 o 仅与 k(g)、k(k)、h 相拼，ê 与其他声母相拼。威妥玛方案另有 ui、uei 两个字形等价于 PY 的 ui, 但 uei 仅与 k'(g)、k'(k) 相拼，ui 与其他声母相拼。威妥玛方案的 uo、o 均等价于 PY 的 uo, 但 uo 仅与 k(g)、k'(k)、h、sh 相拼，o 与其他声母相拼。

由于印刷技术上的考虑，威妥玛方案的某些区别符号可以省略，例如 ê 与 ŭ 上面的帽子一般都能省略，因为这种区别符号省略后不会导致二义性。但是表示送气的符号“，”和 ü 上的两点不宜省略，否则 chu (朱)、ch'u (出)、chü (居)、ch'ü (区)就无法区分了。

下面我们用威妥玛标音方案给下列汉语句子标音(无需标调)：

休息以后，代表们参观了我们的教室、录音室、办公室、图书馆、食堂和礼堂等地方，还了解了我们的学习和生活情况。参观以后，代表们提出了不少宝贵的意见。

Hsiuhsi ihou, taipiaomên ts'ankuanlê womêntê chiaoshih, luyinshih, pankungshih, t'ushukuan, shiht'ang ho lit'ang têng tifang, hai liaochiehlê womêntê hsüehhsi ho shênghuo ch'ingkúang。Ts'ankuan ihou, taipiaomên t'ich'ulê pu shao paokueitê ichien。

3.2.4 法式汉语标音方案　法语国家最早的汉语标音方案是由S.

Couvreur 创制的，曾用来为他主编的《汉语词典》(1890)标音。后来，Arnold J. A. Vissière 在 Couvreur 方案的基础上加以简化和改进，编制了一套自己的汉语标音体系，于1902年发表在《汉语语音的法式标音方法》(Mèthode de transcription française de son chinois) 一文中，并应用于他所著的《汉语入门》(Premières leçons de chinois)(1909) 一书。目前，Vissière 的方案在法国被视为最有权威的汉语标音方案。由于这个方案经常应用于《远东法兰西学院公报》(Bulletin de l'Eoole Française d'Extrême-Orient) 之中，所以有时也称为 BEFEO 方案。

本方案只有六个声母与 PY 一致：f, l, m, n, w, y。不一致的声母有：ng (零声母), p(b), ts'(q 与 c), tch'(ch), t(d) ts(j 与 z), k'(k 与 q), p'(p), j(r), ss (仅等价于 si 的声母), ch(sh), t'(t), s(s 与 x), h(h 与 x), tch(zh), k(g 与 j)。

法式方案有下述韵母与 PY 一致：a, ai, an, ang, ao, ei, eng, ia, ian, iang, iao, ie, in, yu, yuan, yue, yun, ing, iong, o。与 PY 不一致的韵母有：aï(ai), en(en 或 yan 之中的 an), ö(le 之中的 e), o(o, e, uo), e(ê, ye 之中的 e, zhi 之中的 i), ong(ong 和 eng), eul(er), eu(zi, ci, si 之中的 i), i(mi 之中的 i, 不变读), ea(lia 之中的 ia), ien(ian), eang (liang 之中的 iang), eao(liao 之中的 iao), ieou(iu), eou(ou), ou(u), iu (ü 或省略双点的 u), oua=oa(ua), ouai=oai(uai), ouan=oan (uan), iuan(juan 之中的 uan), onang=oang(uang), iue(xue 之中的 ue), ouei=oei(ui), ouen=oen(un), iun(qun 之中的 un), ouo(uo)。

法式标音方案的变读比较多。例如声母 k 兼表 PY 之 j、g；声母 k' 兼表 PY 之q、k；声母 h 兼表 PY 之 x、h；声母 ts 兼表 PY 之j、z；声母 ts' 兼表 PY 之 q、c；声母 s 兼表 PY 之 x、s。也就是说，其中 PY 之 j、q、x 各等价于两种法式拼法，可任选，但它们都必须紧跟着韵母 i，以这种条件

变体解决变读歧义。韵母也有变读情况，例如：o 兼表 PY 之 o、e、uo，e 兼表 PY 之 ê 与 zhi、chi、shi、ri 之中的韵母 i。另外，还有 PY 的一种拼法对应于法式两种拼法的情况，如不少带介母 i 的复韵母在与声母 l 拼写时，都得将 i 改为 e，试比较 niang（娘），leang（凉）。又如 PY 之 e（鹅）对应于法式的 ö 与 o 两种拼法，其中 o 仅与 ng（零声母）、k、k'、h 等声母相拼，ö 与其他声母相拼。再如 PY 之 uo 对应于法式的 o 与 uo 两种拼法，其中 uo 仅与 k、k'、h、ch 相拼，o 与其他声母相拼。法式方案的韵母 ong 也属条件变读，当它的声母为 p'，p，m，f，w 时读作 PY 之 eng，否则仍读 PY 之 ong；另外，PY 之 eng 也对应于法式 eng、ong 两种拼法，但它们是不同的分布变体。在法式方案中，尚有许多自由变体，可以任选，例如 ai=aï，ouan=oan 等等。

这套法式标音方案有声调符，必要时可以采用，其标记是：阴平“-”，阳平“^”，上声“\`”，去声“´”。

下面看看用法式方案如何给下列汉字标音的情况（按音序排列）：

1．事 che；	2．深 chen；	3．耳 eul；
4．否 feou；	5．西 hi；	6．先 hien；
7．会 houei；	8．混 houen；	9．柔 jeou；
10．江 kiang，	11．就 kieou；	12．青 k'ing；
13．君 kiun；	14．个 ko；	15．故 kou；
16．老 lao；	17．落 lo；	18．忙 mang；
19．熬 ngao；	20．朋 p'eng；	21．夏 sia；
22．信 sin；	23．索 so；	24．私 sseu；
25．抄 tch'ao；	26．的 tö；	27．在 tsai；
28．见 tsian；	29．去 ts'iu；	30．娃 wa；
31．夜 ye。		

3.3 俄文字母的转写 俄文使用斯拉夫字母，将俄文字母转写为

拉丁字母是文字罗马化的一个重要组成部分。转写是两套字母表之间的转换，因而是一一对应的。经过转写的文字还可以既容易又准确地恢复原有的字母形式。目前，俄文转写一般采取两种大同小异的方案，它们的区别仅在于某些俄文字母按第一方案应转写为单个的或带附加符号的拉丁字母，而按第二方案则转写为拉丁字母组合。例如 ш 按第一方案转写为 š，按第二方案转写为 sh。在缺少带附加符号的印刷或输入设备的条件下，转写可多采取字母组合来代替那些带附加符的字母。具体转写规则参见表 5，当两种转写法相同时，只注一个，不同时，分别以(1)和(2)注明。

下面给出两个例子，供大家参考。其一是将俄文题录转写为拉丁字母(不同的转写法在括号内列出)，其二是将罗马化的俄文题录恢复成斯拉夫字母的原文形式。

1. Проблемы семантики при автоматическом анализе и переводе.

Problemy semantiki pri avtomatičeskom (avtomaticheskom) analize i perevode (perevodje).

2. Statisticheskaja kharakteristika frazeologicheskikh edinic francuskogo jazyka.

Статистическая характеристика фразеологических единиц французского языка.

在语言自动处理领域中，还可使用一种特殊的俄文转写规则。这种规则既不用附加符号又不用字母组合，而是采用一些数字来转写字母。而且转写有时注重字形而不注重语音（例如俄文的 ь 转写为数字6，俄文的 ю 转写为拉丁字母 н)。同时，下述带下加线的字母转写较为独特，与表 5 的规则不太一致。具体规则如下[①]：

①引自 Bruderer, H.E. Handbuch der maschinellen und maschinenunterstützten Sprachübersetzung, 1978.

表 5　俄文字母转写拉丁字母规则表

俄文字母	拉丁字母
А	A
Б	B
В	V
Г	G
Д	D
Е	(1) E　(2) JE 或 E
Ё	(1) Ë　(2) JO
Ж	(1) Ž　(2) ZH
З	Z
И	I
Й	J
К	K
Л	L
М	M
Н	N
О	O
П	P
Р	R
С	S
Т	T
У	U
Ф	F
Х	(1) H　(2) KH
Ц	C
Ч	(1) Č　(2) CH
Ш	(1) Š　(2) SH
Щ	(1) ŠČ　(2) SHCH 或 SHH
Ъ	(1) ..　(2) ,,
Ы	Y
Ь	,
Э	(1) Ê　(2) EH
Ю	(1) Û　(2) JU
Я	(1) Â　(2) JA

А—А, Б—В, В—V, Г—G, Д—D, Е—Е, Ж—J, З—Z, И—I, Й—1（数字), К—K, Л—L, М—M, Н—N, О—O, П—P, Р—R, С—S, Т—T, У—U, Ф—F, Х—X, Ц—Q, Ч—C, Ш—W, Щ—5（数字)，Ъ—7（数字)。Ы—Y, Ь—6（数字)，Э—3（数字)，Ю—H, Я—4（数字)。

3.4 用俄文字母为汉语标音的方法 用俄文字母为汉语标音问题是由著名汉学家毕丘林（Н. И. Бичурин）于1839年在他的《汉语语法》一书中提出的。他所制定的标音方案经过瓦西里耶夫（В. П. Васильев）的一些修改曾用于1867年出版的一部词典中。后来，又经过一些小的改动，这个方案又用于帕拉基·卡法洛夫(Палладий Кафаров)[①]和波波夫（П. С. Попов)1888年在北京编纂出版的汉俄词典。后来此方案被称为帕拉基方案。作为一个正统的俄文转写汉语的方案，它后来没有再经历什么实质性的修改，一直普遍用于苏联和其他使用斯拉夫字母的国家。

下面用两个表来对比一下汉语拼音和帕拉基方案。

表6 声母对照表

b—б,	p—п,	m—м,	f—ф,
d—д,	t—т,	n—н,	l—л,
g—г,	k—к,	h—х,	j—цз,
q—ц,	x—с,	zh—чж,	ch—ч,
sh—ш,	r—ж,	z—цз,	c—ц,
s—с,	w—в,	y—(参见说明(1))	

（左边是汉语拼音，右边是帕拉基方案——下同）

① Палладий 是教名。卡法洛夫曾随同俄国正教会在北京住了三十年，对中文有很深的造诣。

表 7　韵母对照表

a—a,	ai—aй,	an—aнь,	ang—aн,	ao—ao,	e—э,
ei—эй,	en—энь,	eng—эн,	er—эр,	i(z、c、s 之韵母)—ы,	
i (其它场合)—и,	ia—я,	ian—янь,	iang—ян,		
iao—яо,	ie—e,	in—инь,	ing—ин,	iong—юн,	
iu—ю,	o—o,	ong—yн,	ou—oy,	u—y,	wu—y,
ü (及省略两点之 u)—юй,	ua—ya,	uai—yaй,	uan—yaнь,		
uang—yaн,	üan (及省略两点之 uan)—юaнь,				
ue (及省略两点之 ue)—юэ,	ui—yй,	un—yнь,	uo—o,		
ün (及省略两点之 un)—юнь					

几点说明：

(1)凡汉语拼音以 yi 或 yu 打头的音节，须去掉 y 之后，再以剩余部分在韵母对照表中查找相应的帕拉基方案的韵母。凡汉语拼音以 ya 打头的音节以及 ye 和 yong，，均须把 y 换成 i，再到表中查找。同样，汉语拼音的 you 应以 iu 在表中查找。

(2)在帕拉基方案中，有三个声母在与两类不同的韵母拼读时对应于汉语拼音不同的声母，形成条件变读。具体规则如下：

帕拉基声母	后续韵母[①]	对应的汉语拼音声母
с——	第一类韵母 (e/и/ю/я)	x
	第二类韵母 (a/o/y/ы/э)	s
ц——	第一类韵母	q
	第二类韵母	c
цз——	第一类韵母	j
	第二类韵母	z

(3)汉语助词“的”习惯上标注为 ды。

(4)硬音符 ъ 可用来作隔音符，例如：

фанъaнь=fang'an (方案)。

下面是用帕拉基方案为汉语句子标音的例子：

①这里划分的第一类韵母为细音，亦称齐齿呼与撮口呼；第二类韵母为洪音亦称开口呼与合口呼。

1. 桌子上有三本新书。

Чжоцзы шан ю сань бэнь синь шу.

2. 学习俄文难不难?

Сюэси эвэнь нань бу нань?

3. 你知道这个字吗?

Ни чжидао чжэ гэ цзы ма?

4. 他在对外贸易学院当教员。

Та цзай дуйвай маои сюэюань дан цзяоюань.

3.5 用俄文字母为日语标音的方法 以上介绍了用俄文字母为汉语标音的帕拉基方案，现在再谈谈用俄文字母为日语标音的问题。这里介绍的是 А.Е.Глускина 与 С.Ф.Зарубин 合著的《简明俄日词典》中所采用的标音方案。下面用三个表介绍它的标音规则。表 8 是五十音图标音对照表，表 9 是浊音及半浊音标音对照表，表10是某些片假名标音对照表。

表 8

あ а	い и(й)	う у	え э	お о
か ка	き ки	く ку	け кэ	こ ко
さ са	し си	す су	せ сэ	そ со
た та	ち ти	つ цу	て тэ	と то
な на	に ни	ぬ ну	ね нэ	の но
は ха	ひ хи	ふ фу	へ хэ	ほ хо
ま ма	み ми	む му	め мэ	も мо
や я		ゆ ю		
ら ра	り ри	る ру	れ рэ	ろ ро
わ ва	ゐ и		ゑ э	を о

表 9

が га	ぎ ги	ぐ гу	げ гэ	ご го
ざ дза	じ дзи	ず дзу	ぜ дзэ	ぞ дзо
だ да	ぢ дзи	づ дзу	で дэ	ど до
ば ба	び би	ぶ бу	べ бэ	ぼ бэ
ぱ па	ぴ пи	ぷ пу	ぺ пэ	ぽ по

表10

ブ б	ビ бь	ヴ в	グ г	ド д
ズ з	ク к	ム м	プ п	ス с
ホ х	ツ ц	ヂ дь	ミ мь	ニ нь
ピ пь		ディ ди	ティ ти	シャ ша
シュ щу	ファ фа	シェ ме	ヌイ ны	イエ е
ジ— ж(жи,зи)		ジャ— жа(зя)		
ジュ— жу(зю)		シ— сь(ш)		
ル— л(р)		リ— ль(рь)		
チ— ть(ч)		フ— ф(х,ху)		

几点说明：

（1）う段各行假名的标音有时可省略元音 у，例如：す—с，ふ—ф。

（2）拗音的标音法如下：

きゃ кя　　きゅ кю　　きょ кё

しゃ ся　　しゅ сю　　しょ сё 等等。

（3）在元音字母上添加“-”即可标示长音，例如：きょう кё̄。

（4）促音采用双写其后辅音的办法标注，例如：がっこう— гаккō。

（5）硬音符 ъ 起隔音作用，例如：ぜんや—дзэнъя。

（6）为片假名标音，除了与平假名相同者外，尚有某些独特之处（参见表10），例如允许元音较多地脱落，有一些特殊的假名组合，同一假名可能有不同的标音法（在括号内给出），等等。

下面是用俄文字母为日语标音的例子：

どりょく
1. 努力——дорёку

しょくりょう
2. 食糧——сёкурё̄

えんじょ
3. 援助——эндзё

4. この問題について彼と意見が合わない

— коно мондай-ни цуйтэ карэ-то икэн-га аванай

5. キルギズ・ソヴイエト社会主義共和国

— киргидзу совето сякайсюги кёвакоку.

3.6 译音问题 译音，即用一定的汉字标记外文的实际语音。其主要对象是标记外国人名、地名，有时也用来音译术语。由于使用这一方法的人常受自己方言的影响，不了解该词的实际发音，或选择汉字不同，因而译音词常不统一，形成一种“公害”。

3.6.1 “公害”的种种表现

（1）一名多译。很多人名不止一两个译名，有的人名竟有二十多个不同译名，如 Менделеев 有 28 个译名① 。最近的一个例子更为典型：世界畅销书之一《IACCOCA 自传》在我国一年之内竟然产生了“雅科卡”、“艾科卡”、“艾柯卡”、“亚科卡”四个译名。

（2）异名同译。如 Haley, Halle, Halley, Hallie, Haley 都译为“哈利”。Claus, Clauss, Clouse, Crouse, Klaus, kraus, Krauss 一律译为“克劳斯”② 。人们很难知道汉字所代表的究竟是其中的哪一位。

（3）音节增多。如 Grunsfeld “格伦斯菲尔德”，原文两个音节，拿汉字一译音，变成了六个音节。又如，Gramstorff “格拉姆斯托夫”，原文两个音节，译音后又变成了六个音节，而且还省去了 r 音。

（4）译音不准确。Young，有人为了避免单个汉字而译为“扬格”，无形中增添了“格”音。Johnson 译为“约翰逊”也离原来的音太远。

（5）原形不复存在。经汉字译音后，很难再变回原文。这样，就必须记两套：汉语是哪些汉字，原文又是哪些字母。

---

①张青莲：《关于“门捷列夫”译名的讨论》，《化学通报》1954.4.

②参见辛华编的《英语姓名译名手册》(修订本)，商务印书馆，1983。下同。

(6)性别混乱。按照《汉字译音表》的规定，人名字尾可以分男女，瓦、亚、林、纳等为男性字尾，娃、娅、琳、娜等为女性字尾。然而，在语言学界曾发生过这样一件事：由于对一个苏联语言学家的性别没有摸准，作者的译名上册为“契克巴娃”，下册改为“契科巴瓦”。

(7)不便识读。1987年2月10日《参考消息》一则新闻中有一段话：“政治顾问小米切尔·丹尼尔斯即将离开白宫，去主持印第安纳波利斯的赫德森研究所……”（下面还有一系列人地名，从略）。如果人地名用汉语拼音照抄不译，段落分明，更便于识读，试比较：“政治顾问 Michal · Daniels Jr. 即将离开白宫，去主持 Indianapolis 的 Hudson 研究所……”。

(8)不利于切分。中文信息的计算机处理要求以词为加工单位，没有词的界限的汉字文章必须进行“切词”预处理。象(7)中的前一段，如让机器来切词，至少有两处很难切，一是“小米切尔”会切成“小米”和“切尔”，二是“印第安纳波利斯的赫德森”也不好切。而后一段，即夹用拼音词的一段则不同，全句各词(组)的界限已比较分明。

(9)影响同胞间的交际。如果联系到三胞工作或华语通用区，问题就更大了。由于几个地区译法不同，看报时常发生理解困难。例如，Reagan, 我们译作“里根”，香港译作“列根”，台湾译作“雷根”。又如，Thacher, 我们译作“撒切尔”，香港译作“戴卓尔”。再如，非律宾总统，我们叫阿基诺夫人，新加坡管她叫科拉桑(有时我们用全称，科拉松·阿基诺总统，但用的是“松”，而不是“桑”)。长此下去，定会形成不应有的语言障碍，影响同胞之间的交际。

用汉字译音，有没有好处呢？唯一的好处就是不会汉语拼音或不懂外语的人能够认识，仅此而已。如果说，在术语翻译问题上，主张意译的人还能以一些理由(如增加“察而见意”的效果)来强调使用汉字的话，那么在译写人名问题上，汉字却一点儿也帮

不上忙，因为这里“无意可察”。这就是为什么某些坚决主张术语意译的同志也主张用汉语拼音译写人名的原故[①]。

3.6.2 产生这种公害的原因　我们认为，既有外在原因，又有内在原因。外在原因有三：(1)新的外国人名不断出现，《英语姓名译名手册》之类的参考书根本概括不了。(2)这些新人名如何读音从一般词典中查不到，有的甚至连外国词典也还未收入。这样就给译音造成了困难。(3)时间性很强(如新华社的新闻稿)，来不及仔细推敲和多方查询。内在原因只有一个，那就是汉字根本不适宜于用作译音工具。这是最要害的问题，即使制定了“译音表”也无济于事。公害的种种表现已足以证明这一点。

3.6.3 公害的长期存在　以汉字译音方式来翻译人名，已有两千多年的历史。汉语拼音产生(1958年)之前，虽然有人也感到这是一个问题，但无奈没有一个更好的工具来代替汉字。有人(如陆志韦先生)甚至曾想利用注音符号来代替汉字译音，这颇有攀高(汉语拼音)不成而就其次的意味[②]。那为什么汉语拼音出现之后还依然如故呢？我们认为，一是对这个问题认识不足，没有看到这个公害依靠汉字是消除不了的，而且在当今情报爆炸的年代，不仅消除不了，而且会越来越严重。二是习惯势力太大，没有足够的决心是办不成的。我们对标点符号推行史值得回忆一下。现今的标点符号比古代的“句读”强得多，这是无可否认的事实。但是，为了推行这种新式标点，先辈们冲破了重重障碍才得以成功[③]。近年来，计量单位的推行以及《关于出版物上数字用法的试行规定》都是冲破习惯势力的先例。希望在人、地名转写上也能迈出这样一步。三是语言规范化、标准化工作没有抓紧，对汉语拼音的推广应用没有给予足够的重视，没有很好地利用汉语拼音这个有

---

①例如陶坤同志，他主张意译的文章见《化学物质的中文命名问题》，载《科学通报》3卷6期，主张用汉语拼音转写外国人名的文章见1962年8月22日《光明日报》。

②陆志韦：《外国语人地名统一问题》，《中国语文》1953年8期。

③凌远征：《标点符号推行小史》，《语言教学与研究》1986年第3期。

力工具来解决当今社会上存在的一些语言问题。四是在用汉语拼音转写人地名方面还有些具体问题(如怎样发音的问题)还没有得以很好地解决。

3.6.4 解决问题的途径　处理外国人名、地名的翻译问题基本上有两条途径,一条是采用同音替代译音法,另一条是实现译名国际化。其中第一种办法是在同音字里选出适合音译的汉字，其他的均为替代，以此适当减少用汉字音译所造成的混乱。当然，同音替代法不能治本，而只是一种权宜之计。第二种办法则是建立在汉语拼音化之上的一种治本的办法，就是说用拼音字母转写外国人、地名。虽然具体实施时尚可分层处理，但实现译名国际化应当成为我们的理想。

3.6.4.1 同音替代译音法　同音替代译音法虽然不可能从根本上消除译音的种种公害，但可在一定程度上避免一名多译所造成的混乱，因而不失为一种权宜之计。人们之所以采用同音替代是因为意识到文字只要表音就够了。事实上，文字拼音化就是同音替代的最高发展。

下面向大家介绍的是中国地名委员会制定的英汉和俄汉译音表(见表 11、12)。表中横向为辅音，纵向为元音，交叉处为音译汉字。经过同音替代加工的译音表大大缩减了同音字，使大家在采用汉字译音时有一定之规可循。少数尚有同音字处(在括号内给出)是为了区别人名的性别以及地名的特征。由于英文单词拼写与读音经常不相吻合，所以英汉译音表是以读音而不是以拼写为基础的,采用此表时必须先查得该词的读音(包括国际音标和韦氏音标两种),才能查得译音汉字。至于俄汉译音表，则直接给出原文的拼写形式以及相应的罗马字母转写形式，这样就可以根据原文或转写形式查出译音汉字。

当然，这两个译音表只是试图为以后的译音选取汉字提供某些规范，至于以前已经约定俗成的译名则不宜更改，否则反而增加新的异体。下面给出按译音表查得的译音实例，供参考。

英文译音实例：

Lansing ['lænsiŋ] 兰辛

Travis ['trævis] 特拉维斯

俄文译音实例：

Молдавия　　莫尔达维亚

Ставский　　斯塔夫斯基

3.6.4.2 照抄并附加读音指南　照抄是一种特殊形式的转写。如果照抄英文，实际上是一种直写，因为汉语拼音和英文同是拉丁字母，而且字母表完全一致，根本用不着转写。当然，照抄也包括把德文 β 转写为 ss，同时也包括对非拉丁字母的转写。这种方法最便当，而且最能解决问题。吕叔湘等一些老专家就主张采用这种方法①，我们也赞成这种方法。照抄或转写主要是为了统一，力求形体一致，而不注重发音是否相同。不少学者都曾发表这样的看法，在人名问题上，“(形体)统一比(发音)正确重要”。这可以说是他们的经验之谈。实际上，许多约定俗成的东西都是经过多年的混乱而最后固定在一个名称上的，它不一定比其他的更合理、更科学，而只是出于大多数人习惯了的缘故。照抄或转写的好处就在于能达到这一点，不仅能做到国内统一，而且还能做到国际上统一。例如，我国的四川省过去译成外文时五花八门，英文为 Szechwan，法文为 Setchouan，德文为 Setschuan。近年来，由于ISO(国际标准化组织)提倡转写，而且人们也承认了转写的好处和名从主人的原则，因而目前国际上已一律写成 Sichnan。

至于说读法，这也是要考虑的。对于一般人来说，读得不太准还没关系，如有人把姓 Ōu 的“区”读成 Qū，把姓 Shàn 的“单”读成 Dān (但“乐”字不同，有姓 Lè 的，也有姓 Yuè 的，这要问本人才能确定)，或者把墨翟 (Mò Dí) 念成 Mò Zhái。但是，担负语言规范化任务的人(如播音员)，则应该读得正确，

①见《不如干脆照抄原文》，《北京晚报》1985年4月19日。

为大家树立一个标准。这也就是为什么很多国家都为播音员专门编制正音词典的原因。不可否认，在读法上也会发生一些偏差（这种情况在用汉字译音时也同样会发生），如事先没弄清，而把比利时化学家 Stas 念成了“斯塔斯”（转引自袁翰青的《外国人名地名音译问题》，1985年4月5日《北京日报》）。这也没关系，发现后加以改正就行了，改嘴总比改形体容易。从这一点上说，转写也比汉字译音强，因为后者既要改嘴，又要改形。当然，最好还是尽量避免出偏差。

3.6.4.3 适当修改并附加对音指南　这实质上是一种转写＋标音的办法，也可称作标音式转写法。外国人名在我们的语言中只是一种符号，而由外国科学构成的人名术语，一旦进入了我们的词汇，就成了我们词汇的一部分。因此，应该有不同的对待方法。具体地说，对后者的处理应音形并重。如将发 [k] 的 c 转成 k，将 ph 转成 f，以及将某些辅音从简化或嵌上元音，等等。总之，要使这个外来词语更容易念出来。当然，在不影响识读的情况下还应该允许尽可能保持原形，以便与原来的人名呼应。如写成《Mendelev 周期表》，而不是写成《Mendeleev 周期表》（这里只是举例说明要削去一个e字母，而不是要把已经约定俗成的《门捷列夫周期表》改成上述形式）。又如，写成“Karington 子午线”，而不是写成“Carrington 子午线”。

3.6.4.4 转写和译音并用　目前，为了照顾不熟悉拼音或不懂外语的人，转写和译音并用也是一种办法。1987年2月18日《光明日报》上的一篇文章《朗克逝世一百周年国际学术讨论会简介》就是采用这种方法的。全文两千多字，外国人名十七个，全都是先写汉字译音。然后在括号中照抄原文。虽然多费了一些篇幅，倒也清楚。我们相信，如果这篇报道是发表在学术刊物上，完全可不必使用汉字译音。事实上，现在很多学报都已这样做了。这可以算是一种分层处理吧，即学术刊物上外国人名照抄或转写，一般情报刊上转写和译音并用。而一旦汉语拼音应用进一步扩大，再

完全过渡到只用转写的方法，即5.6.4.2中介绍的方法。

外国人名地名的翻译问题是一个老大难问题。不狠下决心是解决不了的。当然，光有决心，没有一定措施也是不行的。我们认为，扩大汉语拼音应用范围，利用汉语拼音来解决这个老大难问题，是最为有力的措施。如果我们采取了这个措施，可以说，我们就能以最佳方式解决问题，因为汉语拼音使用的拉丁字母是世界上使用最广的一种字母，用它转写或适当作些修改，都能保持与许多文字形体上一致或基本一致，甚至在语音上也有相当大的近似性。

4. 术　语

术语是科学文化发展的产物，科学文化越发达，术语越丰富，术语的制定和规范化的问题也就越来越不容忽视。因此，术语问题在世界许多国家的科技情报界受到普遍重视。术语工作和科技情报工作是密切相关的，科技情报工作越发达，术语也会出现得越多，术语的意译和音译问题也就越来越尖锐。本章准备探讨术语自身的特点和构成规律，术语移植和翻译中的一些问题，并且为术语标准化提出了一个采用汉语拼音的音译方案。

4.1. 术语的定义和特征　术语可以是词，也可以是词组；它们用来正确标记生产技术、科学、社会生活等各种专门领域中的事物、现象、特征、关系和过程。术语具有以下四个基本特征：

(1)专业性。术语是表达各个专业中的特殊概念的，所以通行范围有限，使用人数较少。属于甲行业的术语，从事乙行业的人往往不解其意。这就是人们常说的“隔行如隔山”的含意所在。

(2)科学性。术语的语义范围准确，它不仅是标记一个概念，而且还力求精确，从而能够和其它类似的概念区别开来。一方面是科学产生了相应的术语，另一方面是术语把科学认识的成就固定成语言信息。随着术语本身意义的不断精确化，科学也在不断发展。

(3)单义性。术语在某一特定专业范围内是单义的，这是术语与一般词汇的一个重要的不同之处。例如“运动”这个术语，在物理学中指“物体位置的移动”，在哲学中指“物质存在的形式”，在体育中指“锻炼身体的活动”，在政治生活中指“有组织有目的的

群众性的社会活动。”由此可见，确定专业范围是排除歧义的重要手段。

（4）系统性。在一门科学技术里，术语之间彼此不是孤立存在的。而是有层次，成系统的，也就是说，每个术语的地位只有在该专业的整体概念系统中才能加以规定。这种系统性在一门科学的范畴表或分类表中体现得十分清楚。例如，根据后缀 -ology, -ics, -scope, -tron, -meter, -gram, -graph 等等，可以将术语分为各种层次或不同方面。

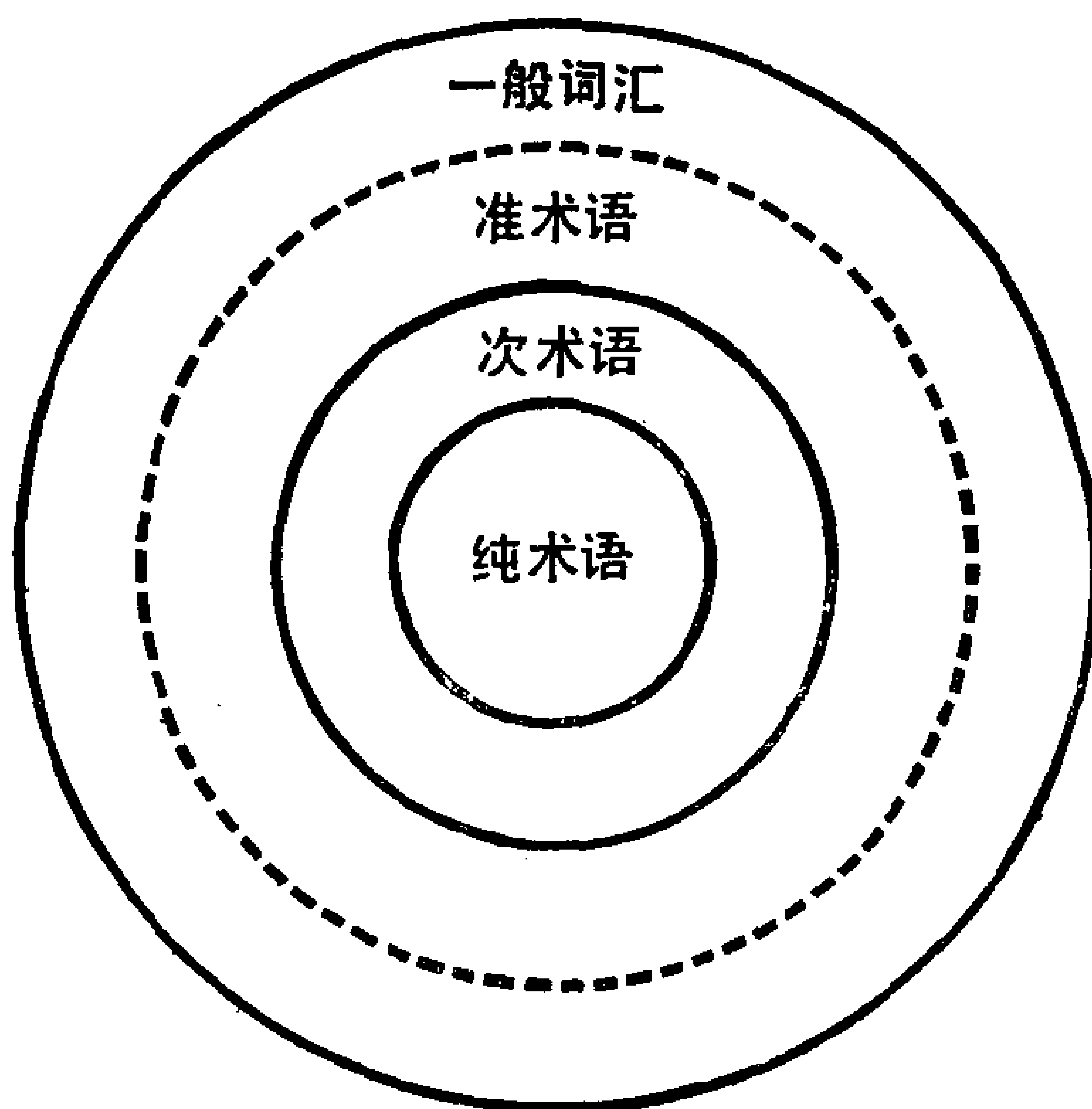
以上四个基本特征是我们用以鉴别术语的重要标志。除此之外，术语一般都不带修辞色彩，且常具有国际性——这两条亦可视为术语的附加标志。

4.2 术语的分类和构成 根据术语的专业学科门类，可将术语分为许多大类，诸如数学类、物理类、化学类、天文类、气象类、地理类、生物类、工业类、农业类、交通类、医药卫生类、军事类、社会科学类，等等。在任何一大类的内部，还可按术语的系统性和语义特征分为若干子类。

术语往往由本民族的一般词汇构成。例如物理学术语中的“疲劳”来源于一般词汇中的并行同音词，然而一旦成为术语后，便与原词的意义发生了不同程度的离异，有时甚至完全失去了联系。

此外，根据术语的使用范围，术语还可以分为纯术语、次术语和准术语。其中，纯术语的专业性最强，次术语次之，而准术语已渗透到人们的生活中，逐渐和一般词汇融合起来。例如：“等离子体”(plasma) 可视为纯术语；“压强”(pressure) 是次术语，“塑料”(plastics) 则只能是准术语了。由此可见，科学技术的普及程度直接影响着术语的三级分类的标准，所以，术语的这种分类只能是一种因时因地而异的相对概念（参见下图）。

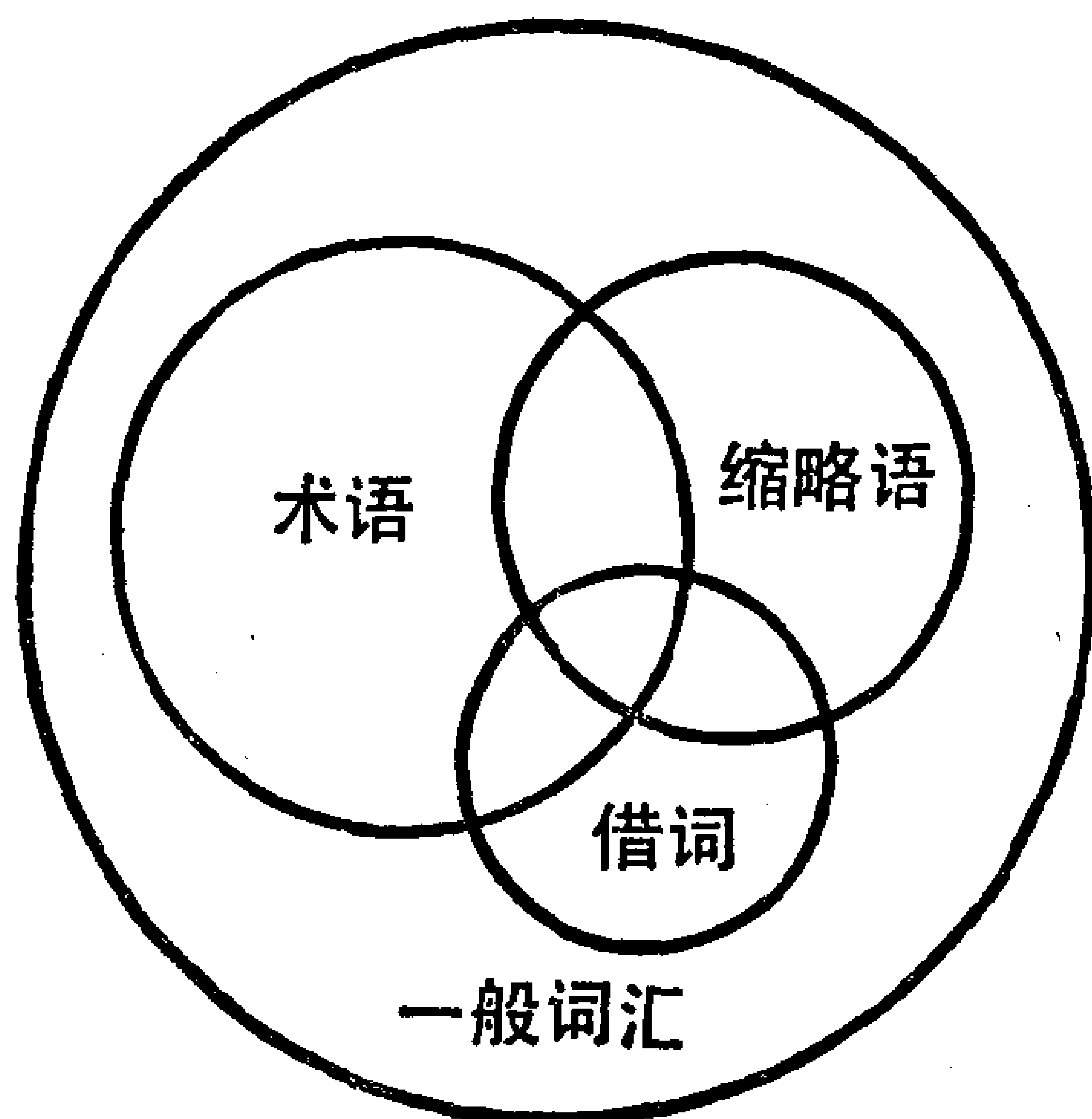
汉语中的术语有不少是外语借词，它们是通过音译、半音译半意译的方式借入的。所谓音译词即是借入语根据借出语的语音转写而成的音近或音同的词。例如：雷达 (radar)、阿托品



(atropine) 等。半音译半意译词是根据借出语的音和借入语的义构成的，例如，"啤酒 (beer)"、"拖拉机 (tractor)"等。虽然术语和外来语借词的引进方式有不少共同点，但是不能在二者之间划上等号。有无专业性是二者的基本区别。带专业性的外语借词才既是术语，又是外来语；无专业性的外语借词只是一般的外来语，如沙发 (sofa)，可可 (cocoa) 等等。

缩略语在具有专业性时也可以成为术语。那些具有专业性同时又是借词的缩略语，称之为外来缩略术语，诸如 "HZ" (hertz) 或"赫"(赫兹)。

术语用借词、缩略语及一般词汇的关系可由下图表示。



4.3 术语的产生、移植和发展 随着科学技术的发展，新事物、新概念不断涌现，人们不得不在自己的语言中利用各种构词手段创造适当的词语来标记它们。由于各国科学技术发展不平衡并且特点各异，不同的术语往往有不同的发源地。例如“训诂”、“叠韵”等术语是我国自创的。重量单位“gramme”（克）是法国自创并首先使用的。

随着文化交流的发展，术语连同它们所标记的新事物、新概念逐渐传播开来，各族人民通过不同的方式把它们移植过来，并使其在自己的语言中生根发芽，这就形成了术语的移植过程。移植方式主要包括意译和音译两种。意译的外来术语，由于获得了译入语的名称，诸如“流水作业”、“激光”、“氢”，一般不易从构词形式上看出它们的异邦渊源。经过音译的外来术语由于借用了术语原来的读音，诸如上面提到的“雷达”、“阿托品”等，在构词形式上有明显的异邦特征。应当说明的是，虽然移植术语的方式有音译、意译两大类，但在我国往往只把音译词或半音译、半意

译词视为外来语或借词[1]，而意译词一旦移植过来，很快就被借入语同化，过不多久人们便忘记它们是外来的了。

在欧洲、美国、苏联，术语的移植一般都采用音译手段，因而术语常常具有国际性。

一般说来，科技术语产生于科学技术发达的国家。同一事物或概念也可能同时在不同国家中出现，因而在不同国度里会产生许多内容相同而形式不同的术语。另外，在移植过程中，不同国家或民族，一个国家的不同地区或一个民族的不同部分，甚至不同的移植者（译者），常以自己惯用的方式移植，这样也产生不少同义不同形的术语。很久以来，术语的产生和移植过程就是这样自发地进行的，因而必然会导致术语的混乱。

4.4汉语在移植外来术语中的问题 术语的混乱现象在汉语中尤为突出。首先，汉语中的术语往往与国际上通用的术语不一致，呈现出明显的“不合群”的倾向[2]；而且异体词很多，内部也不统一。下面举一个例子：

俄　文	гидрохинон	英　文	hydroquinone
德　文	Hydrochinon	法　文	hydroquinone
西班牙文	hidroquinona	捷克文	hydrochinon
波兰文	hydrochinon	芬兰文	hydrokinoni
日　文	ヒドロキノン	世界语	hidrokinono

我们看到，这个术语在国际上流行时基本保持了它的一致性，至多有些微小的变异。在这里，拼音文字间的转写是确保术语实现国际化的基本条件。现在，我们再来看看这个术语在汉语中曾有过多少种名称。

属于音译的有：海德，海得洛，海特路几奴，几奴尼，希特

①从日语借来的词比较特殊，如“场合”、“手续”，是直接从日语汉字借用的，可以称作“形译”。

②据说，这种“不合群”倾向，给学科学技术的外国留学生也带来不少麻烦。他们学了这一套，回去还要学另一套。这或许是到我国来学科学技术的留学生比较少的一个原因。

罗希伦，海得罗几奴，奎诺耳，海德尔，海特困宁，海得儿，海得罗几奴尼；属于意译的有：对苯二酚、氢醌、金鸡纳醇，二羟苯，对二羟苯，对氢醌，对氢化苯醌，对位苯二酚，氢化苯醌，对位·二羟基苯，鸡纳酚，沅化盔农，苯二酚，金鸡纳酚，对位氢醌，对位氢化苯醌，对·二羟基苯，轮二醇，假性二轻氧轮质，困二个醇。

从上面的例证可以看出，音译和意译的任意性是造成汉语术语混乱的基本原因，而且这种任意性因时而异，因地而异，因人而异。为消除术语的混乱，需要语言学家和科技工作者共同努力，解决术语规范化问题。由于国家重视术语统一工作，义同形不同的异体术语并存的五花八门的混乱局面得到了一定程度的整顿。例如上文引述的 hydroqinone 这个化学名词，经过审定，只留下了两个：“氢醌”和“对苯二酚”。由此可见，术语规范化、标准化的问题很重要。名不正则言不顺。术语混乱既影响交际，又影响教育，还影响信息处理。

从国内范围来说，术语标准化是一个国家或一个民族科学技术和文化教育发展的重要条件。这个问题对领土小或分布不广的民族来说还比较容易解决。但是，它对于汉语社会却是一个复杂问题。操汉语的人数量众多，分布面广，并且由于种种原因彼此交往也少，他们在发展各自的文化时都分别吸收了大量术语充实自己的语言。这样一来，不一致的情况时有发生，例如下列一些常见的计算技术名词在中国大陆和海外一些地区就不相同：中国大陆称作“程序”、“硬件”、“终端”、“接口”、“汇编语言”的术语，在海外某些地区被称作“程式”、“硬体”、“端末”、“界面”、“组合语言”。显然，这种情况给汉语社会科技人员之间的学术交流带来许多不便，同时也给青年学生增加不必要的负担。长此以往，不利于保持民族语言的共同性，无益于民族的团结和统一。由此看来，有必要从现在起把加强汉语社会标准化的工作定为语言规划的重要课题之一。

4.5 国际范围的术语现状 从国际范围来说，术语标准化问题则显得更加重要。当今的世界是科技革命的时代。科学技术的发展一日千里，新词术语每时每刻都在增加。例如英汉计算技术词语，1977年版仅有 1.4 万个 1982年版增至大约 4 万个。据有关材料说，目前在发达的语言中90%的词汇是科技术语。鉴于学术交流又是科学发展的必要条件，作为学术交流基本信息的术语如果在世界各国之间不能统一，同义不同形，五花八门的话，无疑会成为影响国际学术交流的障碍。目前，各国的术语有不少已趋于统一，但尚未统一的仍然很多。例如，表达“在单位时间内，为维持一定条件所必须除去的热量”这一概念的术语，英文里有四个义同形不同的术语：cooling lood, heat load, refrigeration duty, refrigeration load; 俄文中有两个：тепловая нагрузка, потребность в холоде；法文中有两个：besoin de froid, demande de froid; 西班牙文有两个：necesidad de frio, carga térmica, 德文只有一个 Kältebedarf。中文译名是“冷负荷”或“热负荷”。类似此种情况比比皆是，不但在各语种之间不能统一，而且在同一语种内部也有异体，这无疑会给学术交流造成不便。

在国际范围内，缩略术语的使用十分广泛。由于科技的发展，统一的缩略语词典已不能满足需要，按专业分科的词典不断出现。应当指出，使用缩略语的原意是节约用语，便于交际，但使用过分，又形成了障碍；因为缩略术语以及缩略专名增多，同形现象也必然增多。虽然分科缩略语词典的出现能够解决一些问题，但同形多义的现象仍然未能消除。

目前，不少国家都致力于术语的标准化。国际标准化组织(ISO) 设有专门的委员会 (TC-37) 从事这项工作，企图促成术语国际化的早日实现。长期以来，各国科技工作者为了掌握世界上先进的科学技术，不得不花费很多时间学习外语和记忆大量不同的科学术语。术语的国际化如能实现，或某种程度地实现，无疑会减轻他们的负担，促进科学技术的更快发展。

4.6 文字体系和术语国际化 术语国际化工作，近年来取得了不少成绩，特别是在使用拉丁字母的语言之间或使用斯拉夫字母的语言之间取得了较好的成绩。但是，非拼音文字体系的术语却难于国际化。特别是汉语，由于采用表意文字(汉字)，过去长期缺乏恰当的书写工具来拼写术语借词。汉字同音的很多，这样使产生了大量异体译名。此外，用汉字书写时没有空格作为词的界限，借词和非借词之间也没有任何标记，这样就增加了词汇混淆的可能性。另外，由于语言结构不同，用汉字音译外文往往造成音节增多，冗长累赘，使用不便。因此，汉语中的国际术语为数有限。日本虽也使用汉字，但同时也使用两套音节字母（平假名和片假名），这样就为外来术语的大量借入提供了方便。据统计，在103个化学元素当中，有81个都采用片假名直接音译外来语，从而使日语化学元素有78%都实现了术语国际化，下面举出一些例子(日英汉对照)：

アルミニウム	Aluminium	铝
カドミウム	Cadmium	镉
カルシウム	Calcium	钙
ガドソニウム	Gadolinium	钆
ガリウム	Gallium	镓
プロトアクチニウム	Protoactinium	镤
プロメチウム	Prometium	钷

通过对照，我们不难看出日文移植术语的方便以及汉字音译的不准确和用字生僻。

另据统计，在日语各专业术语当中，外来语一般都占有半数以上，有时甚至占3/4左右。与这个比例数相比，汉语中的术语借词则显得甚为稀少。试比较下列日英汉术语：

プラスチツク	plastic	塑胶
オゾン	ozone	臭氧
ダイクロイツクミラ	dichroic mirror	二向包镜

ハイドロリツク・トルク・コンバータ	hydraulic torgue converter	液压-扭矩变换器
ドライアイス	dry ice	干冰
エレクトロニクス	electronics	电子学

日语中外来语的数量甚大，但并没有引起日语词汇的混乱，看来也不会引起混乱。因为日语音译的外来语总是以片假名的形式出现的，看上去一目了然，对于词的界限划分也不无益处。鉴于日语和汉语有不少共同之处：它们都是不同于印欧语系的语言，都使用汉字；中文和日文都既能横写，又能竖写，句中无词的界限。日文可以用罗马字标音，中文现在已有汉语拼音，用的都是拉丁字母。考虑到这些共同因素，日语移植术语的做法对于我们来说，有较大的参考价值。如果我们用汉语拼音借用术语，肯定会比日文用假名借用术语更加简便易行，醒目易读。

术语国际化意义重大，起码可以免去"一名之立，旬月踟蹰"之苦，可以减轻青年学生的负担，可以促进学术交流。当然，对于汉语来说，在程度上还会与其他某些语言有所区别，因为汉语历史悠久，在不少方面已经有相当深厚的基础。

4.7 情报工作与术语标准化 术语标准化，情报工作者和技术翻译工作者负有重要的责任。我们认为，必须慎重对待国外科技新术语的翻译问题，不能因人而异。否则，你译成这样，他译成那样，有的音译，有的意译，势必会造成新的混乱，就象 hydroquinone 那样，几十年来被各界人士译成三十多种译名。这样下去，即使有统一的机构不断地审理和统一术语，但仍难以消除与日俱增的新术语译名的混乱局面。因此，术语统一问题和每一个情报、翻译工作者都是密切相关的。当然，术语的统一性和翻译工作的个体性之间是存在矛盾的，解决这一矛盾颇不容易。但是，我们应该从各方面来寻求解决这一矛盾的途径。例如，确定统一的翻译方法无疑有助于术语的统一。

日文里的术语为什么不混乱呢？那是因为他们有一套用片假名转写外来语的转写规则，而这些规则是众所周知的。在我们国家，要使汉语里的术语不混乱，也要有一些原则和规则。否则的话，混乱是不可避免的。为此，对于外文里新出现的术语（这些一般都是纯术语）似乎可以考虑采用音译的方式借入，并且要遵循一定的规范进行音译。

在不得不使用汉字作为音译工具的过渡阶段，我们应当设法克服汉字音译的缺点，特别是不要滥用同音字，而应当严格按照已有的“译音表”选字音译（参见第5章有关译音问题），尽量做到原语的音节与固定的汉字相对应，使得全国各界情报翻译人员有一个统一的音译选字规范。遵守并不断完善汉字译音表是统一汉字借词的一项重大措施。对此，我们必须予以充分重视。音译有规范，混乱才可避免。与此相比较，意译法虽然可以增加一些“熟悉感”或“察而见意”的效果，但是，统一义同形不同的意译术语也十分困难。总之，我们建议对新出现的纯术语最好采用按一定规则进行音译。

4.8 关于术语音译手段的探讨 在我国，音译作为一种翻译手段由来已久，唐代的玄奘法师对有些术语也是主张采用音译的方法。他规定了“五不翻”。“五不翻”的原由包括：秘密故，含多义故，此无故，顺古故，生善故。就是说，在这五种情况下不必翻译（意译），而应当音译。例如其中的“此无故”至今还时常被遵循，亦即对外国的事物采用音译。

谈到音译，无非是沿用汉字或试用汉语拼音，然而无论用哪种文字体制都要遵照一定的音译规则。过去，我们一直都用汉字音译外来术语，结果是站得住脚并在汉语中生根的借词数量很少，不少音译术语都被意译术语取代了。例如“阿屯”（atom）被“原子”取代，“布尔乔亚”（bourgeois）被“资产阶级”取代，等等。正如上文所说，汉语中外语借词很少的根本原因在于文字体制。在使用汉字的情况下，音译法发挥不出自己的优越性。由此可见，

不是汉语不易接受外来语，而是汉字不易接受外来语。如果汉语拼音化实现以后，汉语也完全有可能象其他拼音文字一样，在纯术语的范围内大量吸收音译借词，并逐渐向术语国际化靠拢。

最彻底、最富于改革性的办法就是采用汉语拼音音译外来术语。由于汉语拼音采用拉丁字母，音译过程大多是按照简单的规则进行转写，很容易掌握。用汉语拼音转写的借词不但没有歧义，而且夹杂在汉字文段中使用，一目了然，不会象汉字借词那样，使词汇界限模糊不清。那么，汉字和拼音同文是不是有点不伦不类，使人们感到不习惯呢？我们认为不致于此。而且，习惯是可以逐渐养成的。事实上，在我们的科技文章中，汉字和拉丁字母同文的现象早已屡见不鲜了。例如我们在介绍计算技术的书刊中往往会发现这样一些句子；

在 BASIC 语言中，循环语句由 FOR 和 NEXT 语句构成。

CPU 采用 Intel 8070 A (2MHz), RAM 基本容量 64 KB, ROM 最大容量 384 KB。

另外，近年来在我们的科技论文中，外国人名、地名照写不译的倾向也逐渐普遍起来。

从上面所说的可以看到：中西同文的现象在我们的科技文章中已经不断出现，在汉字行文中夹入一些用汉语拼音转写的纯术语绝不会增加新的生疏感。另据统计，目前日文中罗马字的使用也有增长的趋势，很多术语，尤其是一些缩略语都直接用罗马字母书写。在词典中罗马字的使用更加普遍，排序检索也往往离不开它。例如在日英词典中，就是按罗马字排的。比如“电子计算机”一词在该词典中的格式是：

konpyūtāエンピエークー a computer, an electronic computer.

类似于 konpyūtā 这样的用罗马字书写的术语在日文书刊中也时有所见，从而形成了汉字、罗马字、假名同文现象。

日文能用罗马字拼写外来语，我们为什么不能用汉语拼音拼

写音译术语呢？事实上，用汉语拼音不但可以转写外国的人名和地名(如果外国人名、地名是用拉丁字母写的话，实际上是照抄)，也可以音译国外最新术语。在本章的附录中我们准备向大家介绍一些用汉语拼音音译英语术语的具体规则。假如 hydroquinonl 是新术语的话，就可以译为汉语拼音的 hidrokinon 了(详见附录)。在翻译时，我们遵循的原则是：从字形上说是“名从主人”，这就是说经过音译的术语和原文在词形上仍然保持着相当程度的联系和一致性。然而从借词的读音上说，应当是“主随客便。”例如原文 hydroquinone 读 ['haidroukwi'noun]，但译文 hidrokinon 尽可以读成 [hitəlokinon]，也就是说，可以按照汉语拼音的一般规则读音，必要时注一些附加的规定即可。

4.9 为术语标准化而努力 为了实现汉语的术语标准化，应当建立和健全术语统一委员会。早在建国初期，中央人民政府政务院文化教育委员会曾领导组织了“学术名词统一工作委员会”。近年来，又成立了“全国自然科学名词审定委员会”，负责审定各学科名词的统一名称，并予公布施行。1984年2月27日国家又以命令形式颁布了一批法定计量单位，其中绝大多数都以音译方式定名。例如 mole 原有两个译名：克分子(量)、摩(尔)，现在采用了后者。这样，既免去了记两套术语的麻烦，又达到了向国际化靠拢的目的。

为了促进术语的统一，还应当及时出版各专业的术语词表，使大家随时能够获得一个范本，做到有据可查。在电子信息化的今天，我们特别提倡采用电子计算机编辑出版术语词表，以便能够跟上与日俱增的术语更新的形势。

为了适应情报翻译工作的需要，应当建立术语数据库，从而把术语工作现代化提到日程上来。目前科技文献急剧上升，情报翻译的最大困难即是术语问题，而现有的词典又很难随时更新，跟上技术的发展。如果我们采用计算机把各种专业的术语词典作为不同的文件存储于数据库中，查词典时，只需在终端的键盘上

键入该词，屏幕即可显示该词的意义。由于数据库能及时更新，因而几乎能做到有求必应，因此，采用术语数据库的情报翻译人员能够使自己的产品成倍增加。

问题的关键还不仅是提高翻译速度，更主要的是能提高翻译的质量，促进术语的统一。前几年，小小的 bit 曾有多种译名："比特"、"毕特"、"二进制数"、"二进位数"、"位"等。如果术语数据库很好地建立起来，这类情况显然可以避免。

术语混乱往往是在没有统一规定（即术语词典尚未出版或更新以前）的情况下产生的，而数据库的最大优点是能迅速增补新术语，这样，它便可以起到防止混乱发生的作用。

总之，术语数据库已经显露出它的巨大优越性，它可以为术语工作的现代化铺平道路。近年来，加拿大、美国、西德等国家已建立了几十个术语数据库，并提供咨询服务。其中加拿大的术语数据库 (Termium) 的规模最大，存有 400 万条术语，实际上已成为全国的术语中心。

术语工作的现代化给术语统一创造了有利条件。我们完全可以利用术语数据库来达到术语统一。我们希望成立全国性术语数据库工作委员会，任务是制定术语数据库建立的原则、方法和格式，并监督下属单位遵照执行。委员会之下根据专业分若干工作组，负责建立各科的术语数据库。然后通过计算机网络供各地使用，也可移植到各地。术语数据库的改动、扩充和更新，均应由工作组统一进行。

为了实现术语国际化，应加强标音和转写的研究和各种文字转写方案的设计，建立若干多语种的术语数据库。

新技术革命要求情报工作早日实现现代化。建立术语数据库是情报现代化的一个重要方面。今天，情报界应该而且完全有可能利用术语数据这个新工具，为术语标准化多做贡献。

4.10 术语的标音转写法初探【附：汉语拼音音译规则】 在引进外来术语的问题上，一向有音译和意译的争论。从我国语言的现状

来看，仍以意译为多，即使是少量的音译，也还是采用汉字这一不理想的标音工具。

对于外来语在汉语中已经形成并通用的形式，我们不准备去改变它们，它们已经在我们的语言中生了根，例如：巴黎（Paris），达尔文（Darwin），丙烯腈（acrylonitrile），雷达（radar）等等。但是，对于汉语引入外来语的现代方式，我们是有自己的看法的。术语工作也应该“面向世界，面向未来，面向现代化”。然而汉语在移植术语过程中由于基本上采用了意译法，故与世界上大多数国家的术语形式背道而驰。从这种意义上说，甚至不如我国的一些兄弟民族语言。例如在蒙古、维吾尔、哈萨克等语言中，术语往往采取国际化的词根。面向未来即是向前看。上面说过，已经意译的术语如果已经定型，我们并不想去改变它们。用汉字音译的外来语，即使有种种缺点，但只要是常用的和定型的，我们也仍然承认并沿用它们。但是，出于面向未来的考虑，我们提倡利用标音式转写法来引进外来语和科学名词。这不但是考虑到意译法与国际潮流相违，而且还考虑到汉字作为音译工具是十分不得力的。为了使音译成为引进外来术语的主导方向，就必须采用汉语拼音，这样，我们就可以象日本那样，有可能吸收较多的音译借词了。从某种意义上说，汉语拼音不妨可以成为汉语中的“片假名”。如果有了这种引进外来术语的手段将十分有利于术语工作的现代化，有利于借助计算机建立术语数据库，以及对术语进行翻译、编辑、检索、更新等等。

首先应当说明的是，本章所提倡的术语标音式转写法音译规则有以下三种用途：

一、用来引进国际上最新出现的科学技术名词（纯术语）。对这类术语可直接用本方案书写。由于一经移植的术语已成为汉语中的词汇，因而无特殊必要时不必附原文。

二、用于人名术语的引用。有些科学家的人名已经成为科技术语的组成部分。对这类术语可用本方案为其标音转写，但考虑

到为方便查找该人名在原文中的拼法，应当附加一个原文和本方案拼法的对照表。

三、为各类人名、地名注音。出于名从主人的考虑，人名、地名均应按国际通行的拉丁字母转写法照写原文。然而，为克服按原文拼写的人名、地名不受读音的局限，应在原文后的括号中用本方案为其注音。

为使本方案胜任以上三项任务，它应当具备这样一些特点：

(1)基本沿用汉语拼音方案的拼读规则；

(2)便于拼读，音节结构大致符合汉语的习惯；

(3)标音只求相似，不求准确，关键在于做到便于中国人拼读；

(4)加工的对象以英文为主，兼顾其他采用拉丁字母的语言(非拉丁字母的文字须先按国际通行的拉丁字母转写法转写为拉丁字母，亦可试用本方案加工)，本方案的拼法注意到和原词拼写法保持词形上的相似性；

(5)本方案是深入调查了西方语言(目前以英语为主)的语音和音节构成规律之后，结合汉语的音节特点，在汉语拼音基础上形成的一种标音和转写兼顾的引进外来语的工具。

大致说来，标音方式 (transcription) 考虑语音的因素多一些，为了音似宁可牺牲形似。而转写方式 (transliteration) 则是考虑字母间的对应关系多一些，虽然这种字母间的对应关系也是建立在语音的基础之上，但有时为了形似宁可牺牲音似。举例来说，英文的 computer，引入到德文，成了 Komputer，这就是转写，其中有 c→k 的转化，但是英文的 u 读 [ju:]，德文的 u 读 [u]，这里顾的是形似而牺牲了音似。然而同一个 computer 引入到俄文则成了 компьютер，这种方式恐怕就是标音法，因为对追求音似所下的功夫比转写法要多：其中英文的 u [ju:] 标音为“ю”，这样，语音就较为接近了。

这里提出的汉语拼音音译方案在引进外来术语时力图作到音

形兼顾，标音和转写兼顾。另外，为了适应汉语的读音习惯，在采用汉语拼音音译时也进行了某些音、形上的调整。事实上，其他一些语言在引进外来语时也是采用标音转写法的。例如英文的psychology 引入到印尼文后成为 psikologi。这种音节结构都是根据引入语的音节构成规律而加以调整和变动的，诸如加字母，减字母和改变字母形式等。又如英文的 brandy 引入印尼文为 berandi，加了一个字母；法文的 mètre 引入越南文为 mét，减少了两个字母。其他语言的音译方法对本方案的制定提供了不少启示。我们的拼音音译规则应当具有我们中文的特色，因此，经本方案音译的借词应力求作到易于读音和易于溯源（即是在拼音形式上和外语原文有词形上的联系）。

本着这两个宗旨，下面开始介绍本方案的规则。先从一个药名的音译词说起。英文的 atropine 译为汉语的“阿托品”，我们认为这个译名符合我们的读音习惯，如用本方案标音转化，则为 atopin 不但读音和汉字音译相差无几，且亦能溯源。和原文对比，改动了两处，一是去掉了“r”，二是去掉了末尾的“e”。我们对规则的要求是：标音转写的结果从语音上说应不亚于汉字音译的效果，甚至比汉字音译更加接近原语读音；另外从字形上说可以和原文作到形近，免除了汉字音译的种种不便。

下面按辅音和元音两个部分来讨论这个方案。

一、辅音字母的转写

（1）相同的双辅音字母只保留一个。

（2）c 和 ch 读 [k] 时转写为 k，否则照写。

（3）ck, cq, qu 一律转写为 k。

（4）q 在 e、i 或 y 之前转写为 j，词尾 -age 转写为 -aj，否则照写原形。

（5）几组字母组合的转写：

th→t, ph→f, wh→w, dg→j, sc→s, gh [g]→g, gh [f]→f。

（6）去掉词首不发音的字母：kn→n；ps→s。

（7）x 音标为 ks。

（8）一些固定词尾的转写：

$$\left.\begin{matrix}\text{-tion}\\ \text{-sion}\\ \text{-xion}\end{matrix}\right\}\rightarrow \text{son};\quad \text{-tio}\rightarrow\text{so};\ \text{-nt}\rightarrow\text{n};\ \text{-ce}\rightarrow\text{s};\ \text{-mn}\rightarrow\text{m};$$

$$\text{-mb}\rightarrow\text{m},\ \left.\begin{matrix}\text{-que}\\ \text{-cque}\end{matrix}\right\}\rightarrow\text{k};\ \text{-gue}\rightarrow\text{g};\ \text{拉丁文名词尾}\left.\begin{matrix}\text{us}\\ \text{um}\end{matrix}\right\}\rightarrow\Phi。$$

（9）第二元为 r 的二合辅音连缀 tr, dr, br, gr, pr, fr, cr 等如在词首时照写二元原形，如不在词首则去掉 r，成为 tr→t；dr→d 等等。

（10）其他类型的二合辅音缀如在词尾则去掉第二元，例如：st→s，sp→s，ct→kt→k。如不在词尾则仍照写二元原形，但 sh、ch 不缩略。

（11）第三元为 r 的三合辅音缀无论在何处一律去掉 r，例如：str→st；scr→sc→sk（用本规则及规则 2）但 thr 无论何处都转写为 tr（用规则（5）后成为二合辅音）。

（12）tch、ĉ、cz 均转写为 ch，按单辅音论。shr 亦属二合辅音，适用规则（9）（词尾保留 r）。

（13）经过以上各有关规则的缩略处理之后，如果尚有三合辅音，则在第一个辅音后加衬音"e"，例如：abstract→abstact（规则（11））→abestact（本条规则）→abestakt（规则 2）→abestak（规则10）。

（14）前缀 im、com 当中的 m 转写为 n。

（15）r 在元音之前读音如 l。l 和另一辅音连缀时去掉 l，如 ld→d。其他在元音后面的 r 和 l 均保留，读儿化音。

二、元音字母的转写

（16）单元音一般都照写原形。y 充当元音时转写为 i。元音字母上的区别符号一般只保留"··"和"^"。

（17）二合元音如恰为汉语拼音所有的则一般均照写原文。例

如：ai, ei, ui, ou, ie, ao。另外，ay, ey, ow, wo 也应认为是汉语拼音中所能接受的，必要时将 y 和 w 分别转写为 i 和 u 即可。

(18) 由相同的元音所组成的二合元音，一般只保留一个，例如 oo→o, aa→a, uu→u, 但，ee→ê（注意戴帽）。

(19) 一些汉语拼音中没有的二合元音的转写：

io, eu, ew →iu；ea, ae, eo →ê；oe, oa →o；au, aw →ao；oi, oy →ui。

(20) 三合元音的处理：

a. uai 恰为汉语拼音所有，不变。

b. 去掉最后一个字母，如 aou→ao, uou→uo, uoy→uo, uie→ui。

c. 下列三合元音一律转写为 iu：

(ieu, iew, iou, eou, eau)→iu。

(21) 四合字母 ueue 转写为 iu。字母组合 igh 转写为 ai；ough 转写为 u。

(22) 词尾的 -re 音节去掉 e，元音戴帽"^"。例如：are →âr, ere→êr。

(23) 词尾的 ce 转写为 s。

(24) 其他的词尾字母 e 如在原文中不读音则可去掉。

(25) 经过转写的术语，重音可落在倒数第二音节的元音上，读如阳平。

以上概述了元音、辅音规则共25条。为了便于术语翻译，这些规则可以形式化、程序化，做到用计算机自动转写，一是速度快，二是精确。可以大量地处理外来语的标音转写。我们注意到，实现一个词的转写时往往要调用好几条元音、辅音规则，因此规则序列本身也是规则，而规则软件化正是本系统得以推广的保证。

下面给出几个术语（假设它们是新术语）的标音转写规则的调

用过程，用来说明本方案是如何用于引入外来术语的（括号内的数字是规则的序号）。

1. spectrography→spektogafi

其中：sp→sp (10)

c→k (2)

ktr→kt (11)

gr→g (9)

ph→f (5)

y→i (16)

2. adstrigent→adestinjen

其中：str→st (11)

dst→dest (13)

g→j (4)

nt→n (8)

3. Middlesbrough→Midlesbu

其中：dd→d (1)

dl→dl (10)

sbr→sb (11)

ough→u (21)

4. strathclyde→stateklid

其中：str→st (11)

th→t (5)

c→k (2)

tkl→tekl (13)

y→i (16)

e→Ø (24)

5. hydrocaoutchouc→hidokaochouk

其中：y→i (16)

dr→d (9)

c→k (2)

aou→ao (20)

tch→ch (12)

c→k (2)

熟悉这种标音转写规则之后，就可以较为自如地音译外来术语了。又如：

6. chinophobous→kiunofobous

其中：ch→k，io→iu，ph→f

7. inappreciation→inapeciason

其中：pp→p，pr→p，tion→son

8. microbean→mikobên

其中：c→k，kr→k，ea→ê

9. phthalylsulfathiazolum→ftalilsulfatiazol

其中：ph→f，th→t，um→0

10. Newborough→Niuboru

其中：ew→iu，ough→u

一经标音转写的术语，基本上适合汉语读音。例如上述例10，借助于本方案将原文译为 Niuboru 就好念了，读成近似于“牛伯鲁”的音，这比该词的汉语译名“纽巴勒”的读音更接近原文。总的说来，本方案既保持了可读性，又保持了和原文形态的相似性，因而属于一种标音转写式的音译方案。目前，我们正在进一步完善规则，为本方案的程序化和自动化奠定更加扎实的基础，进而实现术语的计算机拼音音译，达到术语翻译的现代化。

5．机器翻译

翻译和机器翻译是一般与特殊的关系。翻译是两种语言之间在保持意义不变的前提下所进行的语言形式转换，机器翻译也是如此。所不同的只是，机器翻译是借助于电子计算机这一现代化工具来实现这一转换过程的。不言而喻，二者的关系是既相互依赖，又相互有别，因此在下面我们要谈到它们之间的异同。我们拟先从翻译的作用、过程、种类、标准和方法等方面探讨一下有关翻译的原理。在这之后，本章将着重介绍有关机器翻译的过程和方法。首先结合三种外汉机器翻译的流程图介绍机器翻译的全过程。然后分几个专题介绍机器翻译的一些关键问题，其中包括：中介成分体系与动词的配价理论，功能词问题（介词、标点、连词），以及解释程序和表规则算法。

5.1 翻译的原理 翻译的意义在于起一种重要的语言桥梁的作用。世界上的语言千差万别，每种语言都是一种带有某种封闭性的系统。语言障碍至今还在影响着世界各国人民在科学、文化、政治等方面的自由交往。因此，不同语种之间的民族为达到各种交流的目的必须借助于翻译。伴随着交通、通讯手段的日益现代化和世界各国人民交往的日益频繁，翻译的任务大大地加重了。传统的人工翻译已经满足不了形势的需要。所谓"情报爆炸"，即是指情报过剩，其中包括外国文献堆积如山，译不胜译，真正译过来的只是沧海之一粟。在这种情况下，机器翻译可以使这一矛盾有所缓解或部分地解决。由于计算机有着惊人的运算速度，它可以成为翻译的助手，从而能够搭起更多的语言桥梁。机器翻译作

为助手有这样两层含意。一方面，机助翻译作为半自动化的机器翻译可以参与人工翻译的某些过程，例如查找生词、术语、成语和某些句型的用法等等，从而可以有效地提高人工翻译的速度和效率。另一方面，自动化的机器翻译（其中又分全文翻译和题录翻译两种）可以有效地帮助人们获取信息，进行文献初选和评价，还可为进一步的人工翻译拟定一个雏型。在特定领域中的某些成熟的自动化机译系统，甚至可以代替人工翻译。目前，研制机器翻译不但有以上所说的必要性，而且也具备了种种可能性。计算机硬件和软件技术的发展和语言学的深入研究确保了这种可能性。目前在世界许多国家，当然也包括我国，机器翻译正在从可能变为现实。

机器翻译是计算机模拟人工翻译的智能过程，是在程序指令的制导下实现的一种自动化的翻译系统。因此，为了了解机器翻译的过程，首先有必要分解一下人工翻译的各个步骤。我们知道，为了进行语言间的翻译，从事翻译的人必须懂得与此相关的两种语言的词汇和语法规则。如果遇到生词和语法困难，还要查阅词典和语法书。为了使机器进行翻译，也必须具备词典和语法，但这种词典和语法是预先以文件的形式建立在外存贮器（磁盘、磁带等）上的。进行翻译的时候，将有关的文件数据调入内存贮器进行加工。我们知道，计算机的内存贮器相当于人的大脑，因此我们习惯上把计算机叫作电脑。所不同的是，在人的头脑中有一种词汇和语法积累的效应，在翻译过程中已经掌握的词汇和语法就不必再去查字典。而电脑的内存信息是随着一次启动的结束而消失的。所以在机器翻译的过程中对机器词典和机器语法的咨询比人工翻译查词典及语法的次数要高。几乎是逐词查词典，依次查语法。同人工翻译一样，机器翻译也必须根据上下文解决一词多义的问题、惯用语问题、一词多类问题以及对原文进行必要的语法分析。

搞清了原文的词汇和语法之后，下一个翻译步骤即是考虑如何将原文所表达的意义转变成保留同样意义的译文的语言。对于

人工翻译来说，这是一个斟酌的过程。为了实现这个过程，在人们的头脑中必须进行这两种语言的对照、比较，考虑如何生成译文。对于机器翻译来说，这个过程即是原文向译文的转换分析。下一个翻译步骤则是按照译文的语法规则，写出译文。机器翻译也是模拟这个过程，根据译入语的要求组成译文。具体地说即是改变词序，按照新的词序将译入语的词汇检索出来，处理一些前加词或后加词，组成译文。此外，作为翻译的最后一个步骤，还将涉及到译文润色和局部修改，在机器翻译即是所谓修辞加工及输出译文。

综上所述，机器在进行翻译时是模拟人的大脑思维活动的过程，它要经历以下五个步骤，这就是所谓的“翻译五步法”：

(1)查找原文的词义；

(2)分析原文的语法；

(3)进行原文向译文的转换；

(4)按照译文的语法规则生成译文；

(5)译文的修辞加工。

我们注意到，在上面五个步骤中，第一、二两个步骤是处理原文的，第四、五两个步骤是安排译文的，而中间的第三步骤则是解决原文向译文的交替和转换。人们在翻译时自觉或不自觉地都在遵循这些步骤，只是往往处于下意识的状态。而机器翻译正是在分解翻译过程的基础上，在软件指令的指挥下，严格地按照这些步骤进行语言的自动翻译。

5.2 机器翻译的流程介绍 上面我们已经谈到，机器翻译的原理是建立在模拟人的“翻译五步法”的基础之上的。但是作为不同的机器翻译系统来说，还可能有不同的总体设计。最重要的区别在于如何安排原文向译文的转换。转换可以成为一个独立的步骤，这就是独立分析、独立综合系统。转换也可以与原文分析或译文综合同时进行：与原文分析同时进行的称为相关分析独立综合系统；与译文综合同时进行的称为独立分析相关综合系统。本节准

备结合英汉、俄汉、法汉三个系统介绍一下机器翻译的流程图。尽管它们之间在总体设计上有所不同，但它们都属于相关分析独立综合系统。在外汉机器翻译中，这种类型的系统被证明是行之有效的。转换与分析相结合产生了以中介文成分为媒介语的转换分析方式。

这里介绍的三种外汉机译系统都包括四个基本步骤：原文输入（键入或用程序调入原语文句）、原文分析（包括查词典和语法转换分析）、译文综合（按照给定的转换信息调整词序）和译文输出（按照新的序列提取译语词，并按规定的格式显示或打印出译语文句）。下面首先结合外汉机器翻译内的一般设计，简要地介绍一下翻译的整个流程。

5.2.1 原文输入　原文输入可采取不同的方式。原文句子属于数据文件，可采取预制的办法建立原语句子库，存贮在磁盘或磁带上。有的大型机（例如 UNIVAC—1100）还可采取读卡机送入原文句子，建立文件。微型机则可在键盘上直接键入。如果属于随机试验性的，也可以预先不作文件，直接翻译键入的句子。原文句子输入有一些技术性问题须作处理。例如标点符号是作为特殊词汇处理的，其两侧均应留出空格。对有些外语采用附加符号也应作一些人为的规定。这些规定不仅在原文输入中应予贯彻，并且在建立词典和原文分析时也应贯彻。例如在法文中有这样的规定：

É=E2,　Ê=E3,　Ç=C6,　Ë=E5,　Ù=U4

德语中有：Ä=AE,　Ö=OE,　Ü=UE,　β=SS。而且德语凡名词首字母均大写，这是一个十分重要的信息。为了进行区别，可采取：大写字母 A=A$，大写字母 B=B$…等做法。

俄文的原文输入在一般计算机上无法直接实现，通常是将俄文字母先转写为拉丁字母之后再进行键盘输入。例如 д=d，г=g，等等。

鉴于原文人工输入是一个耗费成本而降低翻译速度的工作，

为避免直接输入，可采取有读带机的计算机。读带机可以把磁带中的文句逐句输入，不过，如何进行原语文句的筛选，仍然存在一些必须处理的技术性问题。为实现实用型机器翻译，应当解决好原文自动输入问题。更加理想的原文输入是通过光电阅读装置自动识别出原文字母，并自动转换成机内部的代码。

5.2.2 原文分析 原文分析包括两个阶段：词典分析和语法分析。关于词典分析这里主要讨论综合词典和成语词典。语法分析则是结合三种外汉翻译的流程图分别作一些简明的介绍。

5.2.2.1 词典分析

(1)综合词典的数据结构

机器词典应当包括翻译过程所需的尽量详细的双语词汇的对照信息和可能的转换信息。其中不但应当包括属于原语的静态信息，也应当包括对各种可能的动态用法的模式规约。从功能上说，机器词典应该包括几个分词典，即是：综合词典、成语词典、结构词典以及多义词典。从词典的设置来看，这几个分词典可以放在一起，建立一部包括各种分词典功能的总括性词典，也可以分而设置。下面首先来介绍一下综合词典的数据结构。所谓词典的数据结构即是词典作为数据文件的记录格式，它可以反映词汇初始信息的具体内容。这里介绍的综合词典与其他分词典相互独立，但在综合词典中设立了转分词典加工的键接信息。词典的组织形式可以是相关文件或索引文件。用计算机高级语言编程建立词典时应考虑到词典的易于修改和信息更新。同一般词典一样，机器词典也是按原文的字母顺序排列。查词典时，如属相互文件(有时也叫随机文件)则采取二分法查找较为合理，若是牵引文件，则可用原语词形为键的查找方式。

下面是综合词典的一种数据结构(见下页表1)。词典的数据结构因设计者的构思不同，可以有很大的差异。一个词条记录的信息量应以既充分又不赘长为宜。

查词典之前首先应把原文句子的词切分出来，依次按下标送

表 1　词典的数据结构

词形X(15)	代码	词类	逻辑语义场	形态	词义码	类型特征	配价	配价组合	转换加词	后继
ENDR-OIT	A 1	N	N 9 M A	M	D I F A G					0
LE	A 1	D	D E F N	M S		H	D N			2
LE	H 1	R	P 1 R 3	A K M S	T A		R V			2
REMP-LAC	A 1	V	V 5 G 4 Y 1 E 4	C 1	D A I T I		V N P	P A	2 5	3

入内存开辟的句子场。句子场不但可获得词典信息，还接收削尾所得的形态信息。另外，在语法分析阶段，句子场还不断接收更多的动态信息。削尾图示是查词典程序的必要组成部分。因为英、德、法、俄等外语都有程度不同的词形变化。为了节省存贮，机器词典一般仅存贮词干，其他形态变化均由削尾图示处理。目前，我们的英、法、俄等外语的削尾图示已研制齐全。下面以法语为例，介绍一下削尾图示。

我们对法语词汇的存贮和削尾所持的基本原则是：有形态变化的词大多存贮词根形式，查词时先到词典查询，如果查不到，则削尾后再查。

法语的形态变化主要体现在动词、名词和形容词上，特别是动词，变位系统十分复杂，有大量的词尾变化。考虑到这些特点，在词典中存贮词根形式是完全必要的。例如名词仅存单数形式，形容词和某些有性范畴变化的名词仅存阳性单数形式，其他有关性、数变化则有待削尾解决。动词的词根选择比较复杂，往往要一词存多根。与词根匹配的是一套削尾规则，动词变位的各种形态结论也是靠削尾判定的。

查词时先到词典查询，可以避免无尾可削的词徒然进行削尾加工，从而减少虚功。这样做是因为考虑到法语不变化或不大变化的词在文句中仍占多数。就名词和形容词而言，已经不存在格变化，其单数形式或阳性单数形式往往是秃尾的。就动词而言，虽然变化比较复杂，但对一些使用频率很高的常用词还可考虑存贮它们的各种变形词的形式①。由此可见，采取“先查词典，查不到则削尾再查”的算法是适宜的。

法语动词的变位系统可分为三大组。《法汉词典》（上海译文出版社出版，1982年第一版）还把动词分为70多个细部变位类型，并在词条中以 [C. xx] 的形式标明动词变位表号码。为了

① 有时由于受到削尾规则的制约，某些变形词的存贮甚至是强制性的。

确定各种动词的诸词根形式，研究词尾和词根在接续时发生的形态音位变化，我们从这70多个变位类型中选取了代表性较大的20个作为典型词例，从而能够在一个适中和方便的范围内研究动词的选根和削尾规则。至于其他各种变位类型亦可由此及彼参照处理①。

动词的存贮形式确定之后，就能够统计出每一词根可能后续的全部词尾。对于某一具体的词尾来说。它和某些词根有接续关系，而和另一些词根却没有接续关系。例如词尾 -ons 和词根 fin- 就不存在接续关系。为了把词尾和词根的接续关系搞清楚，有必要把可能出现的变位词尾逐一研究，用形态音位学的观点考查每一词尾究竟能和哪些词根接续。在此基础上就可以编制出削尾规则表。毫无疑问，削尾应按照一定的次序和层次进行。由于削尾的对象包括形容词、名词、动词，且动词又可分为不同的变位组别，因此，同一词尾在不同的条件下也可能具有不同的随机形态参数。所有这些，都必须认真处理。由此可见，编制法语的削尾规则虽然并不困难，但却是一件相当麻烦和细致的工作。

方案的功能在于提供削尾120种，得结论50多条。可以说，只要按照规定的方式存词，查词典程序不但适合任何文体法语文本的词典分析，而且还适用于法语程序辅助教学中的形态分析，如果词典内容扩展，还可用于语词解释和用法说明。

削尾所得的结论虽然有五、六十种，但实际上是二三十种基本参数的不同组配，主要包括有关性、数、人称、时态、语式等形态参数以及词类和成分等语法信息。这些形态信息既可用形态参数表示（每个参数二字节，例如 4S，8P），又可用特征位表示（设置9个特征位，每位占一字节，例如 T4=S，T8=P）。由于每一个形态参数都是由一位数字和一个字母表示的，在使用

---

① 为了适应既定的削尾规则，其他变位类型词应侧重于考虑词典中的存贮方式。

特征位的情况下，前面的数字为特征位号、后面的字母为特征位内容。我们已经做到，任何词尾的各种特征在同一特征上不相交。

基本的形态参数包括以下一些：

1D＝第一人称单数；1S＝第一人称复数；2D＝第二人称单数；2S＝第二人称复数；3D＝第三人称单数；3S＝第三人称复数；4D＝单数；4S＝复数；4B＝第二组变位动词；4Z＝大部分第三组变位动词；4U＝某些具体词根的标记；5M＝阳性；5F＝阴性；6E＝带“E”尾；6Y＝带“ES”尾；6T＝分词特征；7A＝直陈式现在时；7Q＝直陈式未完成过去时；7J＝直陈式简单将来时；7C＝条件式现在时；8P＝直陈式简单过去时；8S＝虚拟式现在时；9X＝虚拟式未完成过去时；9I＝命令式现在时。

削尾可能产生的语法信息有：WY＝谓语；SI＝动词不定式；SA＝形容词；SN＝名词；SG＝现在分词；SE＝过去分词。

根据上面介绍的基本参数，再来看50余条削尾结论也就一目了然了。例如 $_1S_8P$ 表示“直陈式简单过去时复数第一人称”；$_2D_7A_8S$ 表示“直陈式现在时或虚拟式现在时单数第二人称”等等。

作削尾结论时，有时必须查看条件，不同的条件导致不同的结论。常用的条件包括词类和变位组别的信息①。例如“E”尾，其各种结论所依据的条件有动词、形容词（名词）、Y类词，动词内部还可参照变位组别而分为两种情况。

削尾结论表给出了法语形态加工的全部结论。规则用等式表达。等号左边的为结论，右边的为具有该形态特征的各种词尾。圆括号中给出条件参数，说明某词尾仅当符合特定条件时才具有该形态特征。方括号给出加字母信息，说明某词尾被削去之后须再加上给定的字母才可构成词汇的词典形式。

---

① 词类条件包括：名词、形容词、动词、Y类词，分别标记为（＊N）、（＊A）、（＊V）、（＊Y）。属于变位组别4B、4Z、4U的条件分别标记为（＊B）、（＊Z）、（＊U）。当规定的条件并不相符时用（＊--）表示。

(2)成语词典的算法

在机器翻译中，成语的大量存贮对提高译文质量有重要意义。因为很多词组是无法逐词翻译的。也就是说，成语的意义并不以词为单位，而是以词组为单位的。例如英语的 at large, make up to, in the light of 等，根本不能逐词直译。因此必须当作成语整体存贮。所谓成语，是指语言中具有习惯性用法的词组。为了便于查找，在综合词典中应给出成语中心词的特征。例如上述的三个例子 large, make, light 应具有成语中心词的特征。查过综合词典之后按给定的中心词信息到成语词典中查找。成语词典的算法，要从两个方面来探讨，一是造成语词典的算法，二是查成语词典的算法。下面分别加以介绍。

a) 造成语词典的算法

在综合词典中，单词占 15 字节，对于大部分单词来说长度是够用的，如怕不够也可给出 20 字节，但略显浪费。可无论如何单词项长度是定长的。这在综合词典里几乎不成问题。但成语的构成方式纷杂，有两个词组成的成语，有四个词组成的成语，甚至更多，例如法语中“passage de Vénus devant le soleil”（“金星凌日”）为 6 个词组成的成语。除成语构成词数不同，每个词的长度也不同。在这种情况下，如何存贮成语便成为一个必须妥善处理的问题了。

按照机器翻译已往的经验，成语仍采取定长存贮的办法。除成语作为一个义项所应具有的和单词一样的种种信息之外，还应有中心词(15字节)，成语构成词数，中心词的位置，以及包括有30个字节的成语存贮空间。考虑到成语构成词数的多变性，可分为由两个或三个词构成的成语以及由三个以上的词构成的成语两大类。对于前者，每个成语构成词分配给 10 个字节 ($10\times3=30$)。对于后者，每个成语构成词分配给 6 个字节($6\times5=30$)，最多可以不重叠地存贮包括 5 个构成词的成语。当然，如果其中的一个构成词超过 6 个字节，超过的部分则在写入词典时自动削去，

例如 go through fire and water（赴汤蹈火）为由5个词构成的成语，其中第二个词 through 由七个字节组成，存入时，只取 throug 的形式，原字母 h 删略。此外，对于由5个以上的词构成的成语，如上面所举的法语成语为6词成语，passag-e（词尾 e 删略），de，Vénus，devant，le 这5个词已占满30个字节的空间。对 soleil 如何处理呢？我们提出一种叠置法，把第6个词叠置在第1个词上，把第7个词叠置在第2个词上……等等。当然，这种叠置实际上是取代。把 soleil 叠置在首词 passag 上之后，实际存在的只有 soleil 了。因此这个成语的存贮方式如下：中心词 VE2NUS，成语长度6，中心词位置3，成语形式为："SOLEILDE ... VE2NUSDEVANTLE ..."。

以上谈了成语词条存贮的一些特点。同综合词典一样，每一条成语作为一个独立的词汇单位，也具有自己的词典信息和语义参数、配价等词典信息。这些词典信息的设置也同综合词典一样，只要查到了某一条成语，即可调出所需的成语信息。这里附带说一下，在不少情况下，整个成语的词类与中心词的词类是截然不同的，例如 in addition to 整个作为一条成语的词类是介词，但其中心词 addition 的词类是名词。如果某成语中心词可构成一个以上的成语，则按最大匹配原则以构成成语词数的多少依次排列、存贮。

b）查成语词典[①] 的算法

查过综合词典之后，对载有综合词典加工信息的文本句子进行逐词扫描，如果某个词带有构成成语中心词信息，则须转查成语词典。同综合词典一样，成语词典也可采取二分法查找。如采取索引文体的话，成语词典亦可用成语中心词为索引键进行查找。如果成语词典与综合词典的成语信息匹配，文句中带"成语中心词"信息的词一般均可在成语词典中查到。问题在于文句句中的

① 这里介绍的一种成语词典的存法和查法，系 ECMT-78 英汉机器翻译系统，采用"首词表"方法查成语。详见《语言和计算机》第1期第27页。

一些词虽然带有“成语中心词”的信息，但这只提供了一种可能性，在给定的语言环境中，它不一定构成成语。于是在取得成语词典信息之前，有必要对该词在文本中的环境进行是否构成成语的鉴别。首先根据成语词典查得成语构成词数及中心词的位置，然后根据这些信息在文句中取词并构成与成语词典存贮方式相当的格式。这就是查成语词典算法的关键。格式相当是鉴别的前提，如果内容也相当即算查到，可提取成语词典的各种信息，否则不构成成语。鉴别成语首先根据成语构成词数是否少于4而分为两种类型；在大于4时再分是否少于6，以便确定改造为何种格式。例如在实际文句中 adolition 的自然序列号为11，根据查得的中心词位置2和成语词数3，不难算出被鉴别的词为10、11、12号词。如果这三个词均不足10个字节，则由空白补满，否则将多出的字符删掉。在 BASIC 语言中，截取字符可采用 MID 中语句。经字符加工的三个词加在一起即构成与存贮方式相等的格式。如果第10号词为 in，第12号词为 to 当然可鉴别为成语。否则的话，如果第10号词为 the，第12号词为 of，与存贮成语内容不符，鉴别结果为不构成成语。在成语词数大于4而又小于6的情况下，除字符截取方式有所不同之外，其他算法相同。成语构成数大于6的情况下，也必须把文句中的词改造成叠置存贮的形式。

(3)结构、同形、多义词典

同形词典。专门用来区分英语中有语法同形现象的词。例如 close 一词，经过综合词典加工未得到任何具体的词类，而只能得到该词是形/动同形词的指示信息，转到这里后，按照同形词典所提供的检验方法，来确定它在句中到底是用作形容词还是动词。同形词典是根据语言中各类词的形态特征和分布规律构成的。例如在动词、形容词同形的图示中，就有这样的规则：close 带有词尾 er、est 为形容词，处于“冠词+close +名词”和“形容词+ close +名词”等环境时也为形容词，等等。

(分离)结构词典。某些词在语言中与其他词可构成一种可嵌套的固定格式，我们给这类词定名为分离结构词。根据这种固定搭配关系，可以简便而又切实地给出一些词的词义和语法特征(尤其是介词)，从而减轻了语法分析部分的负担。例如，对 effect of ... on (……对……的影响)这一分离结构可给出：

effect＝偏谓，“影响”

of＝前介主 B，“零义”

on＝前介宾 C，“对……的”

多义词典。语言中一词多义现象很普遍。造成多义的原因非常复杂，要知道一个词究竟用作哪个义项，往往难于确定，但这又是必须解决的问题。

多义词的义项选择也主要靠上下文解决。每一个词，不管其本身有几个义项，但在具体上下文中，必然只用于一个义项。为了解决多义词问题，我们必须把源语的各个词划分为一定的类属组。比如说，名词就要细分为专有名词、物体类名词、不可数物质名词、抽象名词、方式方法类名词、时间类名词、地点类名词等等。利用这样的语义类别来区分多义现象，是一种比较普遍的方法。例如 effect 一词，当它前面为专有名词(例如人名)时，要选择“效应”为其词义：Barret effect “巴勒特效应”、Peltier effect “佩尔蒂尔效应”、Joule effect “焦耳效应”等。当它处在表示“过程”意义的动名词之后时就要译为“作用”，如：deoxidizing effect “脱氧作用”、extrusion effect “挤压作用”等。这种利用语义搭配的方法，并非万能，但总还是能解决相当一部分问题。

通过查词典，原文句中的词在语法类别上便可成为单功能的词，在语义上成为单义的词了(某些介词和连词除外)。这样就给下一步语法分析创造了有利条件。

应当说明的是，从结构、同形、多义词典中查找到的信息是同语法分析的内容分不开的，比如说，动词及名词的配价问题实

际上也是结构词问题。事实上，有些系统将这三部分词典加工的内容放到语法分析中去，也同样是可行的。

5.2.2.2 语法分析 这里准备结合俄汉、英汉、法汉三个机器翻译系统的流程图简略地介绍一下各系统的语法分析的内容。从这些流程图中，我们不但可以对各系统的语法分析情况有所了解，还可以看到它们各自的总体设计及整体流程。有关译文综合及译文输出问题虽然也有所涉及，但我们准备在以后的章节中再作具体介绍。

（1）俄汉机器翻译流程图（见下页表2）

机器语法是在机器词典提供的各项条件的基础上进行工作的。它的前半部基本上是一部俄语语法，后半部基本上是一部汉语语法。前半部的任务在于对俄语进行分析，区分同形词尾和确定各种结构成分及其关系。后半部的任务则在于根据汉语规范进行综合，调整各种结构成分的位置，翻译俄语的某些形态成分和修辞。

机器语法是在机器词典提供的各项条件的基础上进行工作的。它的前半部基本上是一部俄语语法，后半部基本上是一部汉语语法。前半部的任务在于对俄语进行分析，区分同形词尾和确定各种结构成分及其关系。后半部的任务则在于根据汉语规范进行综合，调整各种结构成分的位置，翻译俄语的某些形态成分和修辞。

表 2

输入		1 большую	2 роль	3 играют	4 рабочие	5 в	6 деле	7 социалистического	8 строительства.	9
原语分析	词典分析	-ую 单4形， 〈重大〉	-ь 单1.4名， 阴，〈作用〉	-ют 现，复，3 身，谓动， 多义〈起〉	-ие 复1.4名， 〈工人〉	前，多义 〈在…中〉 引状语	-е 单6名， 中〈事业〉	-ого 单2.4形，〈社会主义〉	-а 单2，复1.4名， 中，〈建设〉	
	语法分析	一致定语	单4 直补		复1 主语		状语	一致定语	单2 非一致定语 I(B)	
译语综合	（调整调序主要是）	1_2 $3+(1_2+2)_3$				4_3+3		7_8 $(7_8+8+的)_6+6$ $\{5_6+(7_8+8+的)_6+6\}_3+3$		
输出		$4_3+\{5_6+(7_8+8+的)_6+6\}_3+3+(1_2+2)_3+9$ 工人在社会主义建设的事业中起重大作用。								

这部机器语法的章节如表3所示（根据俄汉机器翻译规则系统第一方案列出）。

表 3

俄语分析	形容词Ⅰ 标点Ⅰ 标点Ⅱ 形容词Ⅱ 公式 标点Ⅲ 动词 名词Ⅰ 名词Ⅱ 句法分析 副句分析
汉语综合	同位语 副词状语 一致定语 非一致定语Ⅰ(A) 非一致定语Ⅰ(B) 非一致定语Ⅱ 非一致定语Ⅲ 后置定语 副动词独立语 插入语 直接补语 间接补语(A) 间接补语(B) 主语 状语 副句 形态加工 修辞

(2)英汉机器翻译流程图

英汉机器翻译 ECMT-78 系统的总体设计如下(表4)：

表 4

原文输入		
原文分析	查词典	成语词典
		综合词典
		结构词典、同形词典、多义词典
	语法分析	动词分析、标点、连词Ⅰ
		名词分析
		标点、连词Ⅱ
		句子分析(包括介词分析)及转换
译文综合	调整词序	加工间接成分及插入句
		加工直接成分
		加工副句
	形态加工和修辞	
译文输出		

下面介绍一下这一系统的有关语法分析的情况(参见表5)。

在词典加工之后，输入句就进入语法分析阶段。语法分析的任务是：(1) 进一步明确某些词的形态特征；(2) 切分句子；(3)找出词与词之间句法上的联系，同时得出英汉语的中介成分。一句话，为下一步译文综合做好充分准备。

根据英汉语对比研究，我们发现，翻译英语句子（其他外语译成汉语，基本上也是如此），除了翻译各个词的意义之外，主要在于调整词序和翻译一些形态成分。为了调整词序，首先必须弄清需要调整什么，即找出调整的对象，根据我们的分析，英语句子一般可以分为这样一些词组(或称词团)：(1)动词词组，(2)名词词组，(3)介词词组，(4) 形容词词组，(5) 分词词组，

表 5　英汉机器翻译过程示例

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
原文输入		The	re-sults	of	such	mea-surements	are	conv-erted	into	the	strength	index	neces-sary	for	eval-uating	the	coking	property	·
原文分析	查词典	冠词, 2.0·	名词, 17.4 ‹25›··	介词, 1.0	形容词, 7.2 ‹28›	名词, 7.2 ‹15›	动词, 1.5 ‹3›	动词, 7.3 ‹6›	介词, 4..8	冠词, 2.0	名词, 3.1 ‹27›	名词, 3.3 ‹10›	形容词, 10.1 ‹18›	介词, 4.4	动词, 7.2 ‹9›	冠词, 2.0	名词, 17.2 ‹5›	名词, 3.1 ‹22›	‹101›
	语法分析						$6_7+7=7$谓语，被动 (削‹3›)								动名Ⅱ，偏谓				
		$1_2+2=2$··· 主A			$4_5+5=5$ 前介定 B‹...110›····					$9_{11}+10_{11}+11=11$ 介宾 A‹121›			前定 B		前介状 B‹116›		$15_{17}+16_{17}+17=17$ 宾 B		
译文综合	调整词序	$(3_5+5‹110›)_2+2=2$							$8_{11}‹121›+11+11$		$12_{11}+11=11$				$(13‹116›+14+17)_{12}+12=12$				
	查汉文词典	‹28›(这样)‹15›(测量)‹110›(的)‹25›(结果)‹ 6 ›(转变)‹112›(为)‹116›(对于)‹ 9 ›(评价)‹ 5 ›(炼焦)‹22›(性能)‹18›(必需的)‹27›(强度)‹10›(指数)‹101›(·)																	
译文输出		这样测量的结果转变为对于评价炼焦性能必需的强度指数。																	

·X.X 是根据该词的语法语义特点划分的小类，例如冠词中的2.0是指定冠词，名词中的3.1是指抽象名词，动词中的7.3是指某种及物动词。

··‹25›为该词的译文代码，即 result 的意义(结果)。这里的数字是假定的。

···$1_2+2=2$表示1已变为2，用同号结成一团，移位时一起移。

····‹···110›：110为“的”字的代码，···表示翻译时“的”字要放在介词词组后。

(6) 不定式词组，(7) 副词词组。这些词组承担着各种句法功能：谓语、主语、宾语、定语、状语等等。其中除谓语外，都可以作为调整的对象。

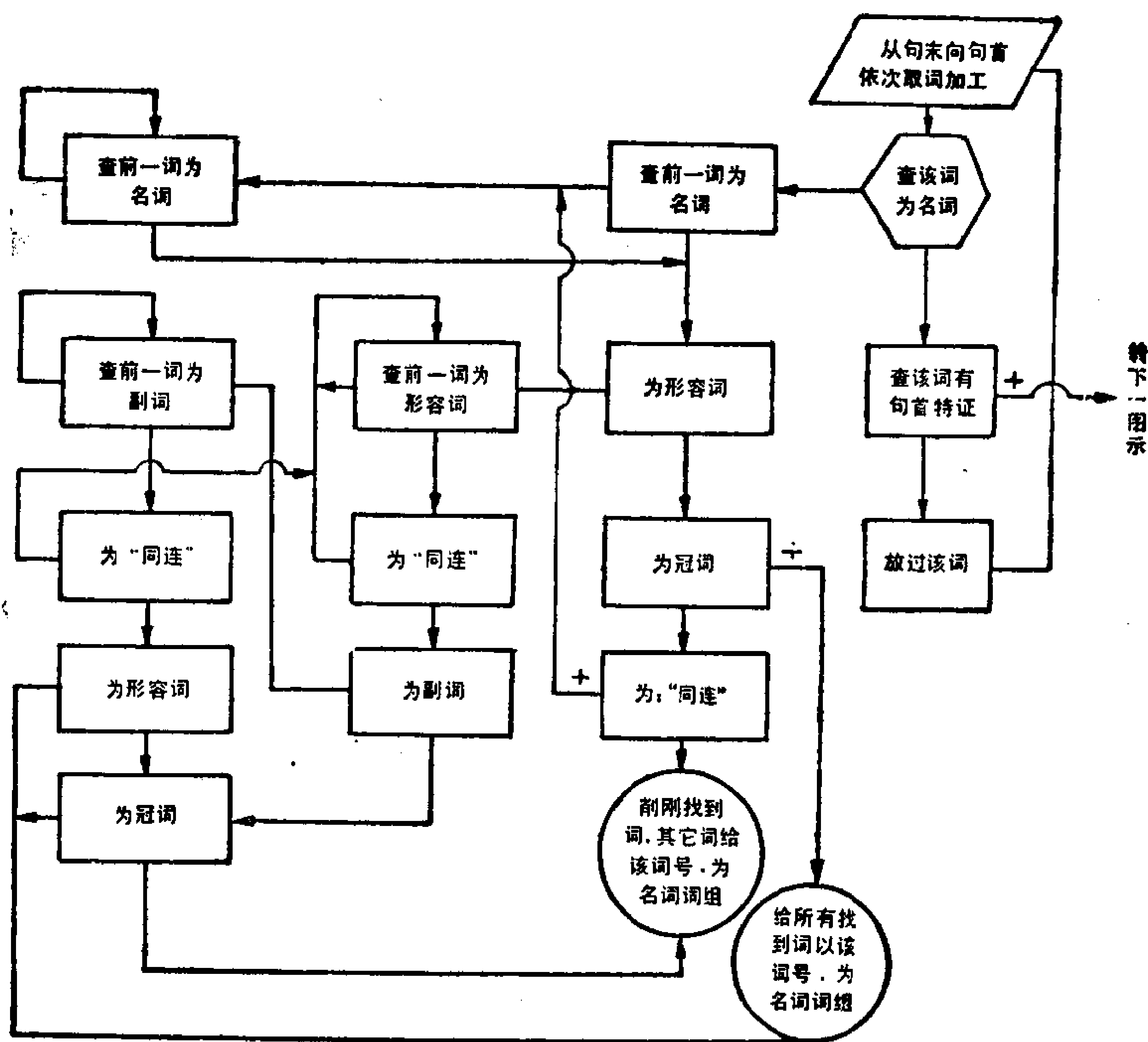
如何把这些词组正确地分析出来，是语法分析部分的一个主要任务。上述几种词组中需要专门处理的，实际上只是动词词组和名词词组。不定式词组和分词词组可以说是动词词组的一部分，可以与动词同时加工。动词前有 to，且又不属于动词词组，一般为不定式词组；-ed 词如不属于动词词组，又不是用作形容词，便是分词词组；-ing 词比较复杂，如不属于动词词组，还可能是某种动名词（如为一般的动名词，则可与其他词构成名词词组），如既不属于动词词组，又不为动名词，则是分词词组。形容词词组确定起来很方便，因为可以构成形容词词组的形容词在词典中已得到"后置形容词"特征。只要这类形容词出现在"名词＋后置形容词＋介词＋名词"这样的结构中，形容词词组便可确定。介词词组更为简单，只要同其后的名词词组连结起来也就构成了。比较麻烦的是名词词组的构成，因为由连词 and 和逗号引起的一系列问题需要解决。

为了简要地说明计算机如何确定词组，下面从翻译规则系统中引出部分框图作为示例（参见下页表 6）。

通过这部分框图，不带连词和逗号的各种名词词组（N; N_1 … Nm; A_1 … Am ＋ N; A_1 … Am ＋ N_1 … Nm; Art ＋ N_1 … Nm; Art ＋ A_1 … Am ＋ N_1 … Nm; Adv ＋ A_1 … Am ＋ N_1 … Nm; Art ＋ Adv ＋ A_1 … Am ＋ N_1 … Nm）便可构成。

词组构成了，调整词序的对象有了，至于是否需要调整以及调整到什么位置，这要根据英汉语的对比特点而定。比如说，英语的后置定语，译成汉语时一般应移到被定语前；英语的排列通常是"动词＋方式状语＋地点状语＋时间状语"，而译成汉语时一般调整为"时间状语＋地点状语＋方式状语＋动词"。

表 6



（3）法汉机器翻译流程图

法汉机器翻译 80 TMFC 试验系统[①] 的各线加工流程图如下（见表 7）：

六五三四

① 参见《语言和计算机》（2）有关法汉机器翻译的介绍。

表 7

<table>
<tr><td>第一阶段</td><td colspan="3">原　文　输　入</td><td></td></tr>
<tr><td rowspan="7">第二阶段</td><td rowspan="7">原文分析</td><td rowspan="3">词典分析</td><td>查综合词典⇄削尾图示</td><td rowspan="3">通过词典分析形成初始符号链</td></tr>
<tr><td>查成语词典</td></tr>
<tr><td>同形、多义、结构词分析</td></tr>
<tr><td rowspan="4">语法分析</td><td>动词、标点、连词分析</td><td rowspan="4">通过语法分析形成中介符号链</td></tr>
<tr><td>名词分析</td></tr>
<tr><td>连词再分析</td></tr>
<tr><td>句法分析</td></tr>
<tr><td rowspan="2">第三阶段</td><td rowspan="2">译文综合</td><td rowspan="2">调整语序</td><td>加工间接成分</td><td rowspan="2">通过译文综合形成终端符号链</td></tr>
<tr><td>加工直接成分和插入句</td></tr>
<tr><td>第四阶段</td><td colspan="3">译　文　输　出</td><td></td></tr>
</table>

结合一个句子的自动翻译的详细过程参见表 8。

下面介绍一下该系统的语法分析的情况。

首先谈谈同形、多义结构词的分析问题。这部分工作就其加工性质而言，属于词典分析的范围，虽然这一系统实际上是采取语法分析的程序加以扫描的，但在这里也应略予介绍。

同形分析的功能在于解决跨类问题。本方案的同形分析是由同形词表和相应的同形子程序构成的。同形词表列出了诸同形词、加工方法及二元结论。这里所谓加工方法，实际上是按照标明的号数转相应的同形子程序进行词类判别。在方案中，7 种同形子程序担负着对下列同形词进行同形判别的任务(见表 9)：

表 8

编　号	1	2	3	4	5	6	7	8	9
原　文	LA	PARTICULARITEJ	LA	PLUS	REMARQUABLE	DE	LA	PLANETE	EST
查综合词典⇄削尾图示	TX R/D	特征 N	TX R/D	DY-CH F.	显著的 A	TX-CH D/P	TX R/D	行星 N	V
查成语词典									
同形、多义、结构词分析	1=D="φ"		3=D="φ"	4 ="最"		6=P	7=D="该"		
初始符号链	{D	N	D	F	A	P	D	N	V
动词、标点、连词分析									9 = GV_1 = 谓语 = "是"
名词分析	$1+2+3+4+5=D_iND_iFA \rightleftarrows D_iD_iFAN=1+3+4+5+2=\underline{2}=GN_1$						$7+8=DN=\underline{8}=GN_2$		
连词再分析									
句法分析与中介符号链	{	GN_1 ② 主 A				P_1 ⑥ 前介的定 B "φ"		GN_3 ⑧	GV ⑨ 谓语
译文综合		$6GP$ + 的 + $GN_1 = 2GN_1$				$P_1 + GN_2 = 6GP$			
终端符号链	{	P_1 GN_2	的	GN_1					GV_1
译　文		该行星	的	最显著的特征					是
备　注	附加符号的变存法：É = EJ，D′ = D″，L′ = L″，È = E； 特征表示：CH = 成语首词，TX = 同形词，DY = 多义词，pl. = 复数，Vant = 现在分词，"φ" = 零词义。 词类表示：N = 名词，A = 形容词，（另见下页备注）								

表 8（续）

编号	10	11	12	13	14	15	16	17	18	19	20	21	22	23
原文	SON	SYSTEME	D″	ANNEAUX	PLATS	L″	ENTOURANT	DANS	LE	PLAN	DE	L″	EJQUATEUR	·
查综合词典 ⇄削尾图示	其 D	系统 · CH N	TX-CH D/P	环 N(pl.)	扁平的 A(pl.)	TX-CH R/D	环绕 V(Vant)	P	TX R/D	平面 N	TX-CH D/P	TX-CH R/D	赤道 N	·
查成语词典		$11+12+13+14=N=\underline{11}$="光环系统"												
同形、多义、结构词分析						15＝R＝"它"			18＝D＝"ϕ"		20＝P	21＝D＝"ϕ"		
初始符号链	D	N				R	V	P	D	N	P	D	N	·}
动词、标点、连词分析		$\underline{11}$＝带前的定 C					16＝现在分词、偏谓语、前的定 C							
名词分析		$10+\underline{11}=DN=\underline{11}=GN_3$								$18+19=DN=\underline{19}=GN_4$			$21+22=DN=\underline{22}=GN_5$	
连词再分析														
句法分析与中介符号链		GN_3 ⑪ 宾A				R ⑮ 后宾B	Vant ⑯ 前的定 C	P_2 ⑰ 前介状 B"在"		GN_4 ⑲	P_3 ⑳ 前介的定 B"ϕ"		GN_5 ㉒	
译文综合		$\underline{16}$Vant＋的＋GN_3＝$\underline{11}GN_3$					$\underline{16}$Vant＋R＝$\underline{16}$Vant $\underline{17}$GP＋Vant＝$\underline{16}$Vant	$P_2+\underline{19}GN_4=\underline{17}GP$			$P_3+GN_5=\underline{20}GP$ $\underline{20}GP$＋的＋GN_4＝$\underline{19}GN_4$			
终端符号链		P_2　P_3　GN_5　的　GN_4　Vant　R　的　GN_3												
译文		在　赤道　的　平面　环绕　它　的　其光环系统。												
备注	R＝代词，D＝限定词，P＝介词，V＝动词，F＝副词，R/D＝代词/限定词同形词；词组表示：GN＝名词词组，GV＝动词词组，GP＝介词词组。													

表 9

编号	代码	类属组	同形类别
1*	D/P	670000	冠介同形
2*	R/D	260000	代冠同形
3*	D/N	610000	冠名同形
4*	R/P	270000	代介同形
5*	V/A	340000	动形同形
6*	A/N	410000	形名同形
7*	V/N	310000	动名同形

同形词表的格式示意如下(见表10)：

表 10

同形词		加工方法	二元结论
certain	3*	(1)1.6.5	(2)6.5.0
des	1*	(1)6.7.4"φ"	(2)7.2.0
en	4*	(1)2.6.2"φ"	(2)7.3.2
l'	2*	(1)2.1.1"它"	(2)6.1.3"φ"
présent	5*	(1)3.8.4(转多义分析)	(2)4.3.2"存在的"
vide	6*	(1)1.5.8"空隙"	(2)4.1.0"空的"
éclat	7*	(1)1.10.8"亮度"	(2)3.15.3"分裂"

本方案规定，类同义不同的词才可称为多义词。那些因词类不同而造成词义不同的词是同形词而不是多义词。

分析多义词，就要调查该词所处的语言环境，然而根据词的语义搭配关系进行多义判别。下面以 plus 为例，说明利用多义判别式分析多义词的过程(见表11)：

表 11

+	−
19/A	20/查△为 plus
A/Ⅰ	B/△$\xleftarrow{0}$6.1.0或6.4.0
B/Ⅱ	C/△$\xrightarrow{1}$9.2.0
C/Ⅲ	Ⅲ/△$\xrightarrow{8}$9.2.0
Ⅰ最	Ⅱ更　　Ⅲ较

例如，当 plus 前面有冠词 la (6.1.0) 时，取 I 词义为“最”。

下面再谈谈结构词分析。结构词，亦即分离结构，从某种意义上讲，是一种开放式的，中间允许插入其他词的“成语”。现以 variation 为例说明分离结构的加工方法：

variation de GN_1 avec GN_2

⟶variation——偏谓语，“变化”

de ——前介主 B，“ϕ”

avec ——前介状 c，“随”

如果句子中 variation 直接后有 de，且 de→有avec，就可以取肯定结论，整个结构译为“GN_1 随 GN_2 变化”，例如：variation de l'éclot avec l'angle de phase（亮度随相位角变化）。

如果 variation 后找不到构成分离结构的介词 de 与 avec，则予以放过。

如果不算同形、多义、结构词分析的话，语法分析则包括四线加工：动词、标点、连词分析，名词分析，连词再分析以及句法分析。

本方案采用“相关分析独立综合”翻译体系，源语向译语的转换分析安排在源语语法的分析之中。在法语和英语这类语言中，词组可以充当调序的基本功能单位。词组的种类有动词词组、分词词组、不定式词组、名词词组、介词词组、形容词词组和副词词组等。我们把实际充当调序功能单位的词组称为“句素”。语法分析的主要任务就在于确定句素，为句素标记中介成分，从而组成中介符号链。

相关分析所涉及的调序加工包括两个方面：一是实施等句素的词序调整；一是筹备超句素的词序调整。

这里解释一下，等句素是指词组内部结构，超句素是指词组外部结构。为了区别这两种不同的结构层次，我们把调序也分为词序调整和语序调整两种。

语法分析各线程序，实际上是诸子程序的调用及不同形式的

组配。常用的语法子程序有：各种放过规则，前、后取词，给同号，移位等。中间结果文件既是输入文件又是输出文件。将一个句子由中间结果文件送句子加工场，实现某线程序对这个句子的扫描加工，然后再将此句送回中间结果文件，如此反复，逐句加工。中间结果文件自从查综合词典而建立起来之后，历经词典分析及语法分析各线，直至译文综合，不断积累或更新着加工信息。如要查看某一线程序的执行情况，只须调打印程序即可。执行句法分析之后调打印程序，不但可观察句法分析的情况，还可获得中介符号链。

法汉机器翻译的语法分析细节详见表8，此处不再赘述。

5.2.3 译文综合 在相关分析、独立综合的系统中，译文综合在语言学规则方面则比较简单，因为有关的转换信息（诸如应调整那些成分及调整到什么地方）在转换分析的过程中已经完成。译文综合的中心任务乃是把应当移位的成分调动一下。因此，译文综合的难点主要是调整词序的程序算法。

如何调动，即采取什么加工方法是一个非常重要的问题。根据层次结构原则，我们认为下面的方法是一种合理的加工方法（不但适用于俄汉机器翻译，同样也适用于英汉机器翻译），这就是：首先加工间接成分，采取从后向前依次取词加工的方法，也就是从句子的最外层向内层加工的方法；其次是加工直接成分，采取依成分取词加工的方法。如果是复句，还要分别情况进行加工。一般复句，在各分句内部各种成分调整之后，各分句都作为一个相对独立的语段处理，采用从句末（即从句点）向前依次选取语段的方法加工；包孕式复句是采用先加工插入句再加工主句的方法，因为插入句如不提前加工，主句中跟它有联系的那个成分一旦移位，它就失去了自己的联系词，整个关系就要混乱。

译文综合的第二个任务是修辞加工，即根据修辞的要求增补或删掉一些词。比如可以根据英语不定冠词、数词与某类名词搭配增补汉语量词“个”、“种”、“本”、“条”、“根”等；再如还可以

根据有 even（甚至）这样的词出现，给谓语前加上"也"字；又如还可以根据主语中有 every（每个）each（每个）all（所有）everybody（每个人）等词，给谓语前加上"都"字等等。

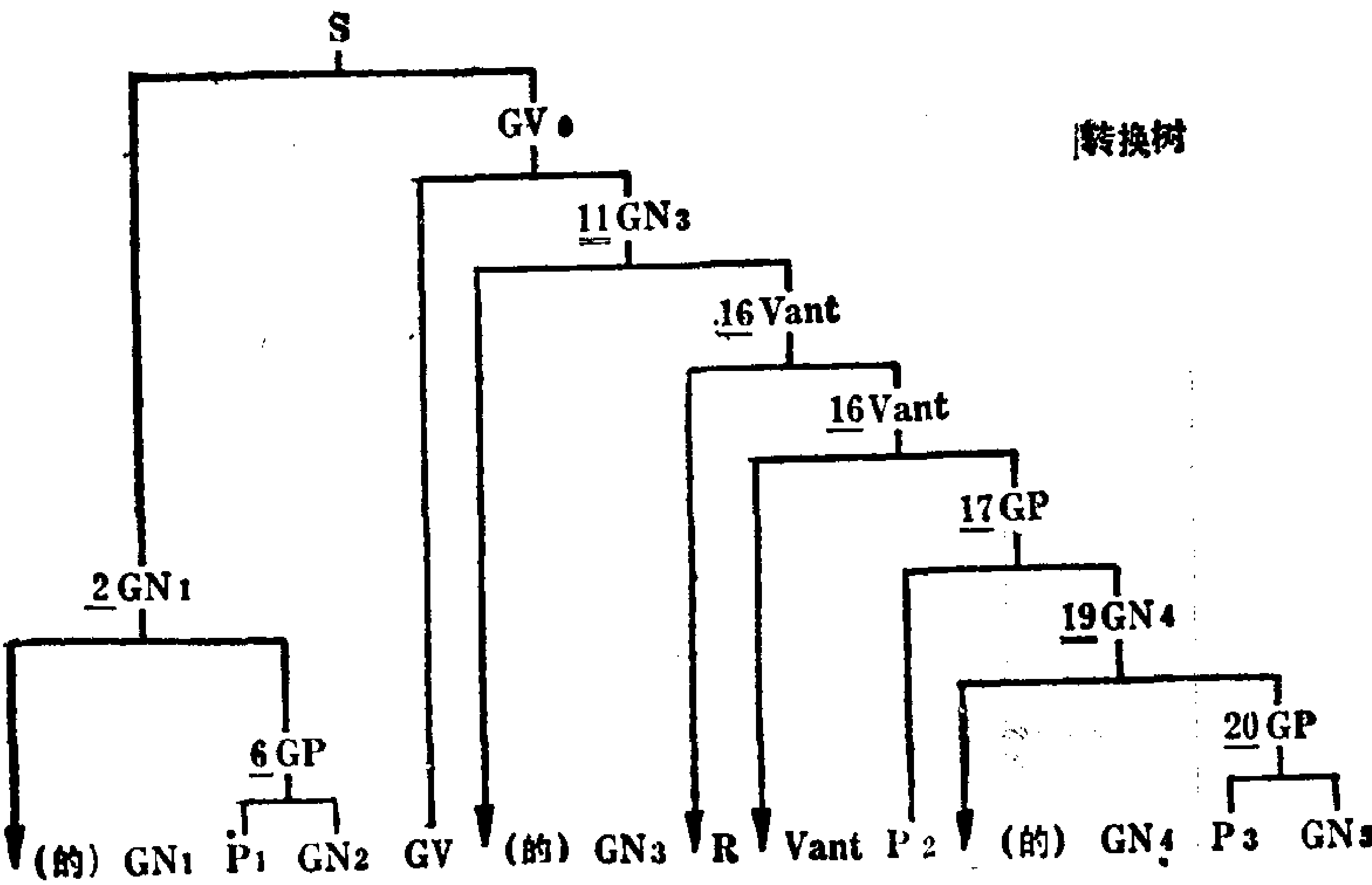
下面以前所引述的那个法汉翻译系统为例，看一看那个法语句子是怎样实现译文综合的。分三个方面来讨论这个问题。

（1）译文综合的任务与转换树

译文综合从句末向句首扫描加工。加工内容主要包括：综合各类间接成分，综合有移位信息的直接成分以及综合插入句。

译文综合是一个按中介成分发出的指令继续给同号以及调整语序的过程。为了反映这种过程，我们以中介符号链为叶造转换树（见表 12）。

表 12



将转换树与示范例句的流程图对照参阅，即可看到译文综合的实施过程。

（2）终端符号链与生成树

在完成译文综合之后调打印程序，即可看到中间结果文件的最终状态。此时示范句已经紧缩为一条仅包括三种词号的紧缩符号链：

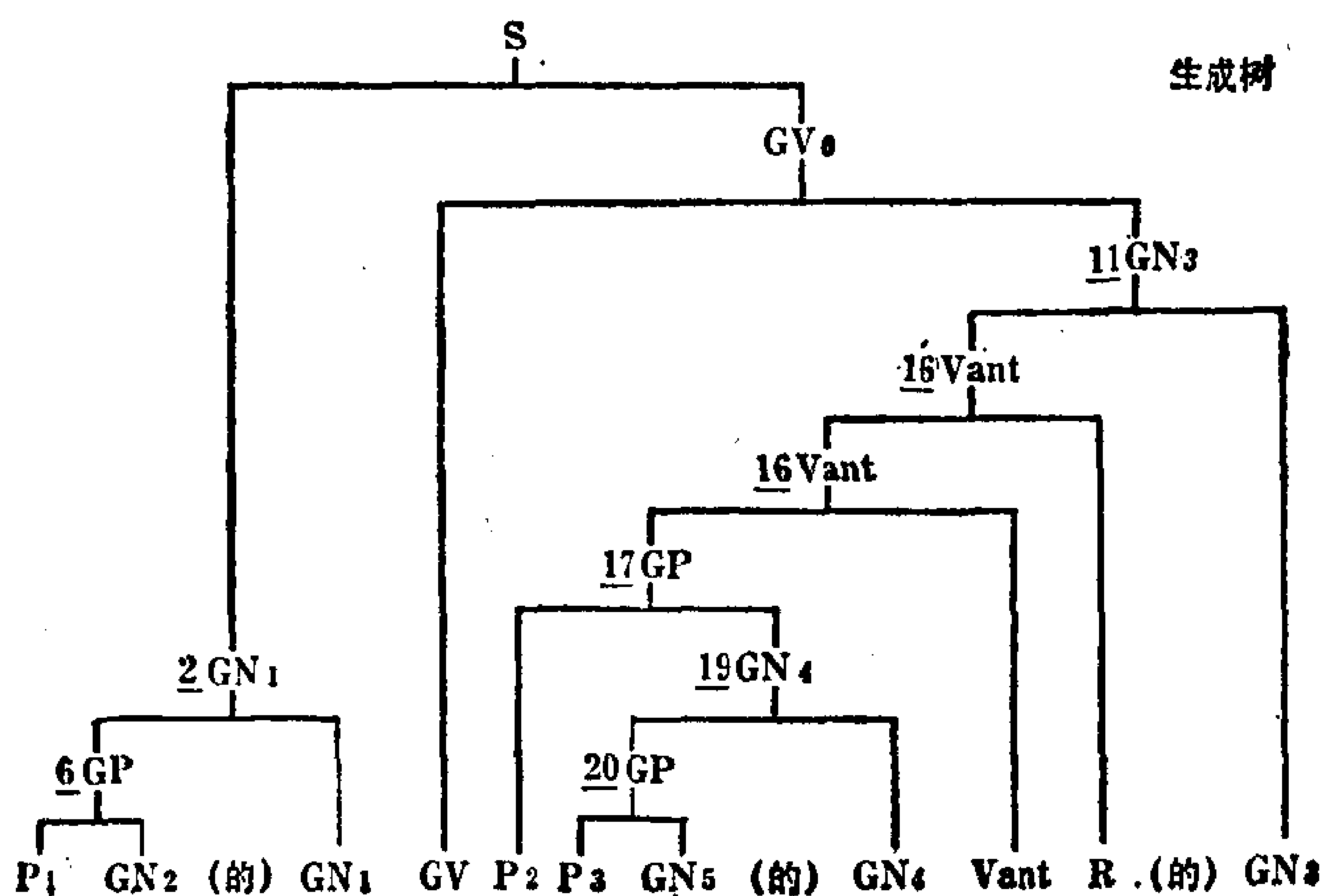
$$\begin{cases} \text{主 A} & \text{谓语} & \text{宾 A} \\ 2GN_1 & GV & 11GN_3 \end{cases}$$

很明显，这条紧缩符号链的主语及宾语均各自包括一群同号词。为了更清晰地观察上述紧缩符号链的主 A 与宾 A 所包含的实际内容，我们不妨仍以中介符号链的句素来表达紧缩符号链，从而形成终端符号链。按照转换树的箭头指向调整语序即可得到终端符号链：

$\{P_1\ GN_2$（的）$GN_1\ GV\ P_2\ P_3\ GN_5$（的）$GN_4\ Vart\ R$（的）$GN_3\}$

以终端符号链为叶造树，即可得到生成树（见表13）。生成树的枝干再现了译文综合的内容，结点多为超句素单位。

表 13



实际上，很多中介成分都可以用语法重写规则表示，这在生成树中体现得尤为明显。现举几例加以说明：

前介定 B：GP+GN⇒GN

前介状 B：GP+Vant⇒Vant

状 B：F+A⇒A

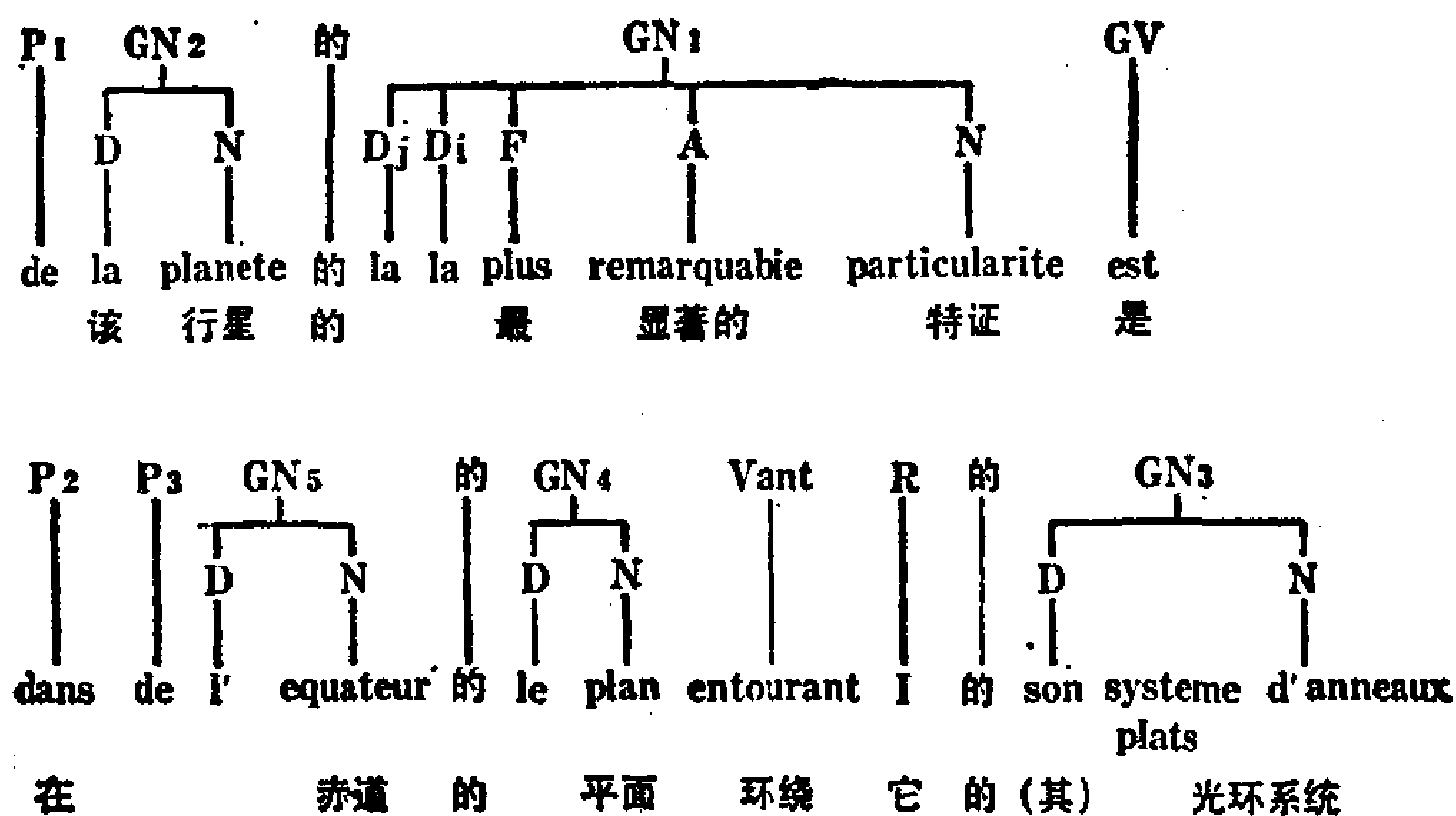
前状 A：F+GV⇒GV

插入句：S′+GN⇒GN

（3）“开花”的生成树与译文生成

终端符号链的诸句素在中介符号链中均曾出现过，考虑到这些句素实际上充当着语义输出单位，所以在翻译的最后阶段，应再现它们在中介符号链中所达到的同号指标下的内部结构（即生成句素的词），这样，就可以使生成树的叶“开花”，形成“开花”的生成树。现将开花部分补充如下（见表14）：

表 14



在“开花”的生成树中，唯独成语不再开花。这时，每一朵“花”代表一个词或一条成语。只要将这些“花朵”依次译出，即可生成汉语译文。

词序调整的实现纯属一个软件技术问题。下面用 BASIC 语

言简明地描述一下调序的基本算法。

首先应当指出，所谓调序是在句子的内存加工场中改变用下标指示的单元地址的内容。当参与量是两个直接单元时，借助于SWAP 语句即可实现两个相关单元的内容互易。但是在实际情况下，参与调序的往往是由相同的序号所表示的一组词 (A)，而它们的调至方位完全可能是跨越一定空间的另一组同号词 (B) 之前。在这种情况下，有必要根据这两组同号词的各自的同号域扫描出它们的起始与结束的单元地址号，将有待于调序的词组 A 送入另一暂设的加工栈，并记载其单元数。将 B 组词与 A 组词之前的所有词按此单元数为增量下移至新的单元，最后再将A组词从栈中取出，送入 B 组词之前空出来的单元，从而实现调序。中介成分里的“前介状 A”即是采取这个算法调序的，其中 A 组词是介词短语，B 组词是动词短语。译文综合的结果是将 A 移至 B 的前边。

事实上，这一算法对“前介定 C”、“前介定 B”、“后宾 B”等都可同样运用。尽管不同的中介成分有不同的调序细节，但总的调序思想仍然大同小异。例如“后宾 B”，虽然属于向后移位，但只要由主程序控制一下两个相关词组的分配，亦可调用以上所介绍的那个算法为基础的调序子程序。

5.2.4 译文输出 译文输出是一个检索的过程。主要是从内存句子加工场中逐单元地将汉语的词义部分取出，即可组成汉语译文。所谓逐单元取汉语词义即是依照句子场下标依次提取。由于译文综合已实现了调序调整，此时句子场中的自然序列恰好符合汉语的句法结构，而原文的序列已被完全打乱了。至于汉语修辞的加词内容一般都是在句子场中的加词项中提取的。如果汉语词义是代码形式，那么还需查汉语词典，转化成汉字，如果已经是汉字，直接调出即可。根据汉语无词汇界限的习惯，还需把多余的空格删去。如果由于条件的限制，汉语词义是按汉语拼音拼写的，那么译文输出时还应确保词汇之间留一个空格。按照目前的

做法，由于有汉字操作系统CCDOS，在造综合词典时即可直接按编码将汉字输入词典，在词典及句子场中，汉语词义一般留4个字的空间，亦即8个字节。如果用代码空间可以少些，如果用汉语拼音，当然就要多些。

译文输出可以翻译一句输出一句译文，也可以翻译一批句子之后输出一批译文。译文即可在终端屏幕上显示，只可在打印机上输出，一般采取外汉对照的形式。

5.3 中介成分体系与动词配价(valency)理论 上面谈了机器翻译的各主要阶段的流程。本节主要就谓语中心论与动词配价问题作专题讨论。

在机器翻译中，我们一向实行以动词为中心的析句法。我们认为，谓语动词是句中的核心，非谓语动词(ving, ved, vint)是句中的次核心。这样做的根据是什么呢？

(1)句子的结构层次是以动词谓语为轴心来划分的；

(2)它是句中最大的联系中心，句中除定语外，其它成分的功能多取决于它；

(3)副句除定语副句外，也同主句谓语有关，作它的主语、宾语、状语；

(4)无主句、命令句一般都由它表达；

(5)许多语法范畴(数、时、态、式)由它表现；

(6)支配关系多由它体现；

(7)语义联系也多由它产生；

(8)动词谓语还是分句中的一个天然界限，利用它可以求得分析工作顺利进行。

动词身上带有那么多有用的信息，分析句子不求助于它又求谁呢？这方面的问题过去已经谈得很多。这里主要是为了再行强调一下。它是语言工程中的一个带根本性的问题。西方的配价理论也主要是以它为基础而发展起来的。

下面首先介绍一下集中体现谓语中心论的中介成分体系，然

后再谈谈西方的一些有关动词配价的理论。

5.3.1 中介成分体系

(1)什么是中介成分体系

中介成分体系是根据外汉机器翻译特点建立起来的一套特殊的成分体系，其中各个成分既不是原语成分，也不是译语成分，而是介于原语和译语之间的句子成分。中介成分是通过原语语法分析和语义分析并考虑到同译语的转换而得出的，它们属于“深层结构”的范围。中介成分不仅能表示成分的功能意义，而且还能表示成分的分布关系(成分与成分之间的联系，成分在句中的位置)和反映两种语言的对比差异。

中介成分体系不单纯是一套标记句子成分的符号，更重要的是解决多对一翻译的得力工具。另外，在它身上体现了我们翻译规则系统的编制方法，体现了我们关于句子切分的理论认识，同时也体现了我们对翻译技术的处理原则。

(2)中介成分体系和翻译系统的关系

大家知道，机器进行翻译，包括三个过程：分析 (Analysis)，转换 (Transformation) 和综合 (Synthesis)。这三个过程可以用三种方法实现：

a) A→TS b) AT→S c) A→T→S

利用第一种方法编制的规则系统，称之为独立分析相关综合的系统。所谓独立分析，就是原语分析只根据原语本身的结构特点进行，不考虑译语的因素，因此分析的最终结果是原语句子成分；所谓相关综合，就是根据原语分析所得的成分进行原语向译语的转换；并在此基础上按译语规范组成译文。

利用第二种方法编制的规则系统，称之为相关分析独立综合的系统。这里的相关分析，是指在原语结构分析的基础上实现原语句子成分向译语的转换，同时在转换过程中要参照译语规范，因此分析的最终结果是一套中介成分；这里的独立综合，是指只需根据中介成分本身提供的移位信息组成译文。

利用第三种方法编制的规则系统，称之为媒介语型系统。这种系统的特点是原语分析和译语综合都独立进行，而原语向译语的转换则利用一个独立的转换程序，即媒介语来实现。

三种类型的系统都可进行一对一的翻译，其中媒介语型系统还可进行逆向翻译。在参与翻译的语言增多的情况下，第一种系统适用于“一对多”，第二种适用于“多对一”，第三种适用于“多对多”。从我们国家来说，如果把汉语译成多种外语(例如英、法、德、俄、日)，采用第一种系统比较适当。这时我们只需研制一套为各语言通用的分析程序(独立的)和五套与之相应的综合程序(相关的)就行了。但是，如果目的是把多种外语译成汉语，采用第二种是完全必要的。这时只需编制五套分析程序(相关的)和一套综合程序(独立的)就够了。如果考虑多对多翻译，例如参加共同体的翻译合作，那当然要依靠第三种，即媒介语型系统。使用媒介语好处有二：一是可减少翻译系统的数目，二是能实现逆向翻译。但是，制定出对多种语言都适用的媒介语，不是一项简单的任务，有许多问题尚待解决。

根据我国目前的具体情况来说，当务之急是将各国的科技文献翻译成汉语。因此，采用相关分析独立综合系统对我解决多对一翻译是最为有利的。而采用这种系统，必然要有一套符号来标记相关分析的结果。这就是我们这套中介成分体系的由来。

(3)制定中介成分的原则

人所皆知，一般划分句子成分主要是根据逻辑语义原则。但是，这样划分出来的成分不能满足机器翻译的需要，因此它们只能表示成分的功能意义，而不能表示成分的分布关系和反映两种语言的对比差异。然而，后者对机器翻译来说，是极端重要的。道理很简单，一个成分，如果本身不能表示分布关系，机器便不能得到它在句中与哪个词发生联系和处于什么位置的任何信息，如果本身不能反映出两种对比语言的差异，机器便不能在综合加工时直接得到把原语改造成译语的任何根据。因此我们结合机器

翻译的特点和要求提出了新的划分成分的原则，并根据这些原则建立起了中介成分体系。

这套成分体系所依据的原则有三：

第一，逻辑语义原则。这一点同传统语法中划分句子成分的原则基本一致。根据这一原则，划分出六大类句子成分，即主语类、谓语类、宾语类、补语类、定语类、状语类。

如果一个成分就其与句子的结构关系来说，是插入性的或独立性的，给以“插入成分”或“独立成分”特征，作为状语的两个次类。

我们的系统是以结构分析为主要手段以语义分析为辅助手段建立起来的系统，因此得出这类成分是很自然、很便当的。这样一些类别，再加上其子类，足以反映语言的各种内在联系。

第二，结构层次原则。句子是分层次的，如上所述，我们主张动词（谓语）中心论，因此认为句中最主要的一层是谓语。谓语是句中最大的联系中心，同时也是调整词序的主轴心。直接与谓语发生关系的成分叫直接成分（简称 A 成分），与谓语以外任一其他成分发生关系的成分，叫间接成分。而间接成分又根据它的联系词的远近分作两种：是前一词或后一词的称为近间接成分（简称 B 成分），不是前一词或后一词的称为远间接成分（简称 C 成分）。这样，上述六大类又可分为以下一些子类：

a. 主语类：主语 A、主语 B、主语 C。

b. 谓语类

c. 宾语类：宾语 A、宾语 B、宾语 C。

d. 补语类：补语 A、补语 B、补语 C。

e. 定语类：定语 B、定语 C（没有 A 成分）。

f. 状语类：状语 A、状语 B、状语C。

另外，根据动词中心论的观点，非限定动词（分词、不定式、以及部分动名词）也是句中较大的联系中心（它们可能带一堆宾语和状语），同时也是调序过程中比较重要的次轴心。因此，在我们的

系统中赋予“偏谓语”特征。“偏谓语”本身可以出现在直接成分层，也可以出现在间接成分层，但它所带的成分则只能是间接成分。

第三，对比差异原则。这个原则要求一个成分能反映出对比语言之间的差异。根据这一原则，句子成分又可以分作以下两类：

a 根据该成分在对比的两种语言中的位置相同与否，可分为“前置成分”、“原位成分”和“后置成分”。

b 据据该成分在对比的两种语言中是否有介词，可分为“介词性成分”和“非介词性成分”。

这套成分体系包括几十个中介成分。应该强调指出，这只是一个基础，人们在处理各自的语言时可以而且必须根据具体情况有所增减，因为任何一对语言都不可能包括所有这些成分，同时也多少存在一些个别现象。实践证明，这个体系不仅解决了相关分析独立综合的问题(即不管翻译什么语言，原文分析是什么样，而译文综合就使用一套)，而且为译文综合创造了极为有利的条件。例如，机器见到“宾A”，便可知道它是同谓语发生联系的直接成分，没有移位信号，故可置之不理。又如见到“前介定B”，便可知道它是一个近间接成分，是介词短语作定语用，有前移信号，而且是要移到前一词(成分)之前，因为B成分的联系词是前一词或后一词。

(4)中介成分如何给出？

要说明中介成分是如何给出的，就要把整个系统的前半部，即分析和转换部分，都介绍一下才能讲得清楚，因为前半部各种操作的总目的就是要得出这种中介成分，为综合汉语译文准备条件。

先讲语法分析。

语法分析可分形态分析和句法分析两方面：形态可提供谓语等信息，例如英语的 takes 和 wrote，俄语的 беру 和 писал，等等。

同形词(即一词多类的词)是一个大问题，这在形态较少的英

语中尤其严重。区分同形词的办法主要有二：一是靠其本身的形态，二是靠上下文。例如 work 是名动同形，如果它在句中出现的形式为 worked，则足能定为动词，如果它在句中出现在 the 或 interesting 之后，则可肯定为名词。词类明确以后，才可谈及句法功能。比如 work 为名词，才能与冠词 the 组合为名词组，然后根据其分布情况，才能分析出它是主语 (the work is interesting) 或是宾语 (I like the work)，或是介词宾语 (I got a lot of experience in the work) 同时又与介词组成短语作状语。

词组或短语是句子结构中的重要一环（词→词组→分句→句子）。词组一般是（当然某些单个的词也可以是）完整概念的负荷者，是句法单位的体现者，同时也是我们调整词序的对象，换句话说，即中介成分。根据英汉对比研究，我们划分出七类词组：即名词词组、动词词组、形容词词组、分词词组、不定式词组、介词词组、副词词组。这些词组，通过上述制定中介成分三原则的加工，便可得出我们的中介成分。比如上面举出的 the work 根据逻辑语义原则区分为主语或宾语，然后可以根据层次结构原则分析成“主 A”或“宾 A”，根据层次结构和对比差异原则，in the work 可以分析成“前介状 A”，即其功能为状语，其层次为 A成分，其构造有介词，其译法要移到谓语前。

在语法分析过程中，介词值得一提，介词作为一种虚词，实际上的作用并不亚于实词。实词虚词的划分，主要是根据能否表达实体意义。一般说来，人们对于实体意义的理解并不难，而介词的结构意义、关系意义纵横交错，却是不易掌握的。介词不仅是指示句中实词之间有某种关系的词，而且借助本身的意义，进一步指示和确立这些关系的内容和性质。如果没有介词，或介词的等价词，人们也就只能表达一些简单的概念。介词在句中起着重要的媒介作用，许许多多的介词又组成各种各样的短语，它们的句法作用不尽相同，它们的语义又丰富多样，因此，介词的处理，对于中介成分的确定，是一个关键问题。

再讲语义分析。

语义分析可以用来精确地描述句法结构、确定词汇单位的意义、说明语义结构与句法结构的相互作用。当单靠词的形态特征和结构特征不足以确定词与词的搭配关系以及词组或句子的结构类型时，语义分析就更显得重要。

例如，下面两句话：

1) We shall have a symposium on Monday.

2) We shall have a symposium on mathematics tomorrow.

结构完全相同，如果单靠语法分析，很难得出 on-phrase 的不同作用。然而，借助语义分析，第一个 on-phrase 中的 Monday 为时间名词，与 symposium 无语义联系，故可判定为状语，作"前介状 A"，第二个 on-phrase 中的 mathematics 为学科名词，与 symposium 有语义联系，故可判定为定语，作"前介定 B"。上文已经指出，介词对中介成分的确定起关键的作用，由此可见一斑。另外，在介词的分析中，不仅能确定句法功能，而且还能确定结构层次和移位与否的信息，同时还能给出相应的词义①。

语义分析，同语法分析一样，只能在一种可靠的基础上才能进行。什么是可靠的基础呢？这就是词的合理分类。词的合理分类之所以重要，是因为它给提供最基本的初始信息。没有这样的信息，任何分析都无从谈起。目前，我们采用的是根据形态、结构和语义三原则制定的一种"类属组三级分类法"。关于语义，多说几句。这里的语义不是指每个词的具体词义，而是一种范畴意义，即从一群词中概括出来的共同特性，例如，day, night, yesterday, tomorrow, week, month, year, hour, minute, second 属于时间名词，从语义场理论来说，这些词同属一个语义场。

语义分析也不是万能的，比如遇到 He saw the girl with

---

①详见刘涌泉、姜一平：«机器翻译与介词研究»，«科技情报工作»，1981. 5.

binoculars（“他看见了带望远镜的女孩”，或“他用望远镜看见了女孩”）这样的句子，就会发生困难。这时，除依靠一个句子内部的语义分析外，还需要依靠更大范围的上下文，或语用学的知识。

（5）实例简介

英译汉的中介成分：

Exchang of thoughts is a constant and vital necessity, for without it,

主A　前介定B　谓　宾A　非同连　介状A

it is impossible to coordinate the joint actions of people in the struggle against

谓　宾A　主A(偏谓)　宾B　前介定B　前介状C　前

the forces of nature, in the struggle to produce the necessary material values;

介定B　前介定B　前介状C　前定B(偏谓)　宾B

without it, it is impossible to ensure the success of society's productive activity,

介状A　谓　宾A　主A(偏谓)　宾B　前介定B

and, hence, the very existence of social production becomes impossible.

非同连　主A　前介定B　谓　宾A

机器译文：思想交流是经常和极端必要的，因为没有它，不可能在对抗自然的力量的斗争中，在为生产必需的物质财富的斗争中协调人们的共同的活动；没有它，不可能保证社会的生产活动的成功，因此，社会生产的本身存在变成不可能。

几点说明：

1．词（词组）下向带＝者属于谓语层，带一者属直接成分层，带～者属间接成分层。

2．necessity 按一般语法书分析是“表语”，机器翻译中定为“宾语”。

3．“for” “；” “and, hence” 都表示分句界限，所以这个句子实际上由四个分句组成，每个分句是机器翻译的一个加工单位。

4．it 这种形式主语在这里不作成分。

5．组成汉语的方法：（1）先加工 B、C 成分，从后向前加

工，定语性的以其前的名词组为轴心，带"前移"信号的需移至找到的名词组前。状语性的以其前的"偏谓"为轴心，带"前移"信号的需移至找到的"偏谓"前。(2)后加工 A 成分，加工 A 成分从前向后，以谓语为轴心。

法译汉的中介成分：

L'échange des idées est une nécessité constante et vitale, car il serait
主A 前介定B 谓 宾A 非同连主A 谓
impossible autrement d'organiser l'action commune des hommes dans la
宾A 前状A 介主A(偏谓) 宾B 前介定B 前介状
lutte contre les forces de la nature, dans la lutte pour la production des
C 前介定B 前介定B 前介状C 前介定B(偏谓)
biens matériels nécessaires, il serait impossible de réaliser des progrès
介宾B 主A 谓 宾A 介主A(偏谓) 宾B
dans l'activité productrice de la société, impossible, par conséquent,
前介状C 前介定B分句点 宾A 前首状A
qu' existât même la production sociale.
非同连 谓 后定B 主A

机器译文：思想交流是经常和极端必要的，因为否则不可能在对抗自然的力量的斗争中，在为生产必需的物质财富的斗争中，协调人们的共同的活动，不可能在社会的生产活动中获得成功，因此，不可能，存在社会生产本身。

德译汉的中介成分：

Der Gedankenaustausch ist eins ständige und lebenswichtige Notwendi-
主A 谓 宾A
gkeit, da es ohne ihn nicht moglich ist, ein gemeinsames Handeln der
非同连 主A 介状A 后宾A 谓 后宾C 前
Menschen im Kampf gegen die Naturkräfte, im Kampf für die Erzeug-
定B 介状C 前介定B 介状C 前介定B(偏谓)
ung der notwendigen materiellen Güter zuwege zu bringen, da es ohne
补B 介主A(偏谓) 非同连 主A
inn nicht möglich ist, in der produktionstätigkeit der Gesellschaft Erfolge
介状A 后宾A 谓 介状C 前定B 后宾B
zu erzielen, und folglich das Bestehen einer gesellschaftlichen Produktion
介主A(偏谓) 非同连 主A 前定B

selbst nicht möglich ist.

后宾A 谓

机器译文：思想交流是经常和极端必要的，因为没有它不可能，在对抗自然力的斗争中，在为生产必需的物质财富的斗争中，实现(协调)共同的行动，因为没有它不可能，在社会的生产活动中获得成功，因此，社会生产的存在本身不可能。

几点说明：

1. 由于德语的谓语、偏谓常后置，所以其状语、宾语常在其前面。原文分析和译文综合时都应照顾到此情况。

2. 根据上面谈的，加工 Handeln (后宾 c)和 Kampf (介状 c)时应向后找"偏谓"，带"后移"特征的应移置"偏谓"后。

俄译汉的中介成分：

Обмен мыслями является постоянной и жизненной необходимостью,

主A 补B 谓 补A

так как без него невозможно наладить совместные действия людей в

非同连 介状A 谓 宾A 前定B 前

борьбе с силами природы, в борьбе за производство необходимых мате-

介状A 前介定B 前定B 前介状A 前介定B(偏谓)

риальных благ, невозможно добиться успехов в производственной де-

补B 分句点 谓 补A 前介状A

ятельности общества, стало быть, невозможно само существование

前定B 分句点 状A 谓 主A

общественного производства.

前定B

机器译文：交流思想是经常和极端必要的，因为没有它不可能在与自然的力量的斗争中，在为生产必需的物质财富的斗争中协调人们的共同活动，不可能在社会的生产活动中获得成功，——因此，社会生产的本身存在不可能。

关于补语A的说明：在直接成分层，有时直接宾语和间接宾语同时出现，为了区分，把不带介词并由间接格表示的宾语，定作补语。至于加工方法，如宾语、补语不同时存在于一层，可以按

同一种方法加工，如同时存在，则按另一种方法加工。

5.3.2 动词的配价理论 上一节介绍了有关中介成分的理论。迄今为止，我国各种外汉机器翻译方案绝大多数都是以中介成分体系进行句法的转换分析的。近年来，我们还注意到国外一些有关动词配价的理论。所谓配价，即是动词对其他词的固有的可预测的支配值，也反映在谓语动词对它两侧的其他成分的强支配功能。一般地说，由于中介成分理论是针对于外汉机器翻译的转换机制而设计的，因此对语言结构所作的动态，转换分析，可能要比下面介绍的国外有关动词配价理论更能直接地应用于机器翻译。此外，国外制定的动补结构式一般只限于对原语进行单一层次的无标记、无转换的独立分析，而中介成分体系不但为结构式标明了成分，划分了层次，还为汉语生成提供了有关移位和加词等转换信息。在这种情况下，我们有必要仍然本着制定中介成分的三原则（逻辑语义原则、结构层次原则以及对比差异原则），来学习和借鉴国外配价理论有关原语分析的研究成果。这样做对我们的机器翻译的研制工作将会大有裨益。本着这一宗旨，我们在分析有关动词配价理论时，常与中介成分理论作对照，并为一些例句标注了中介成分。

根据动词的配价理论，可认为核心句是由谓语及受谓语强支配作用的各直接成分组成的。虽然直接成分都是谓语的直接关联词，但谓语对它们的支配作用有强、弱之分。强支配作用是可以预测的，它反映在动词的配价上面，属于动词固有的属性。而弱支配作用是无法预测的，受弱支配的成分具有随机附加性，不受配价的制约。由此可见，所谓强支配、弱支配完全是从配价角度划分的。在机器翻译中，直接成分无论受强支配还是受弱支配，它们在调序方面的可能性和必要性都是平等的。下面这个德语句子中的“前介状 A”属于受弱支配的成分，但同样需要调序：

Gestern besuchte mein Sohn mit einem Freunde das Museum.

状A　二价Sn,Sa　前主A　前介状A　宾A

（昨天我儿子同一个朋友参观了博物馆。）

谓语的强支配作用体现在动词的配价之上。这些受强支配的成分可统称为补足语。核心句即是一个动补结构式。核心句是最高层次的动补结构式，非限定动词也可构成较低层次的动补结构式。不过本节介绍的动词配价理论所涉及的动补结构式一般都只局限于研究谓语动词。

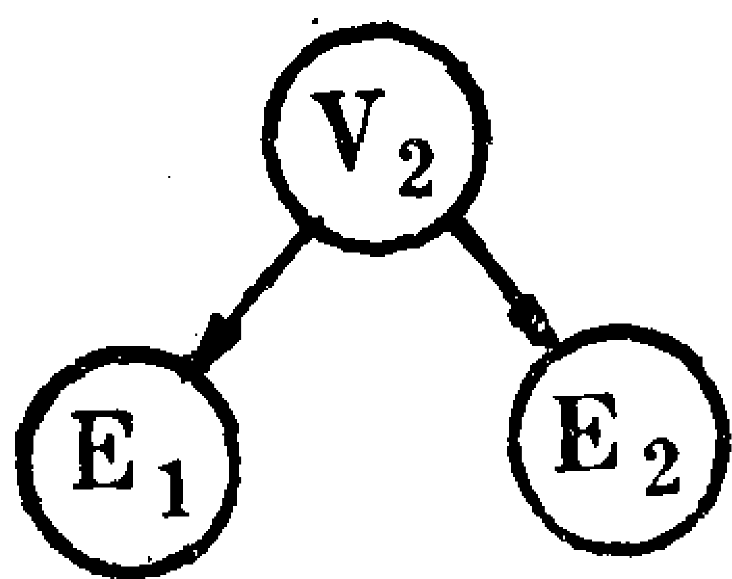
补足语可能包括作为直接成分的主语、宾语、表语、状语和补语（在本节中，补足语和补语是作为两个不同的概念使用的）。它们可能是体词性结构的名词、代词、介词短语、形容词、副词，也可能是述词性结构的不定式、分词和从句。谓语动词的配价对补足语的数量及分布状态有直接的影响，而讨论配价问题必须从研究动词的性质着手。

动词有及物与不及物之分。有的是无人称动词，有的是自反动词。有的要求外位式，有的要求分离式。有的是直接及物，有的是间接及物。有的要求一个后位补足语，有的要求两个甚至三个后位补足语。有不少动词对介词和格范畴都有强支配功能，组成固定框架结构。如果从逻辑语义方面划分、有的属于使役动词，有的属于体态动词，有的要求递系结构。不少动词对它们的主、宾、状语还有语义方面的具体要求。动词的这些特征往往交织在一起，形成复杂的配价模式。

国外语言学界对动词的配价进行过许多研究，下面仅就其中几种有关的理论体系进行一番评介和探讨，最后再同国内提出的谓语中心论作一个比较。

(1) Herringer 与 Helbig 对动词的配价分析

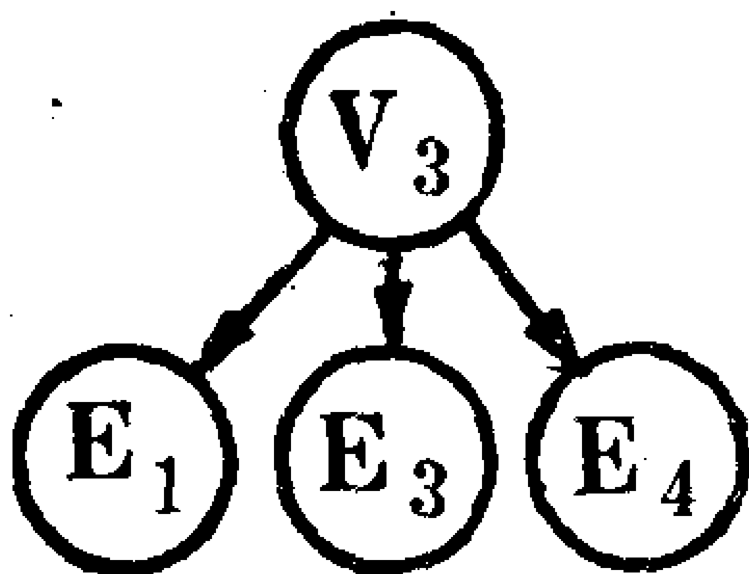
德国语言学家 H. J. Herringer 按照名词的格和介词短语将补足语 (Erganzung→E) 分为五种 E_1、E_2、E_3、E_4 分别表示第一、二、三、四格的名词性补足语，而 E_5 则表示由介词短语充当的补足语。从下面标注有配价的核心句中，我们可以看到：谓语动词的配价决定补足语的数量和性质。



Der Kranke bedarf der Hilfe.

E_1 二价动词 E_2

（病人需要帮助。）



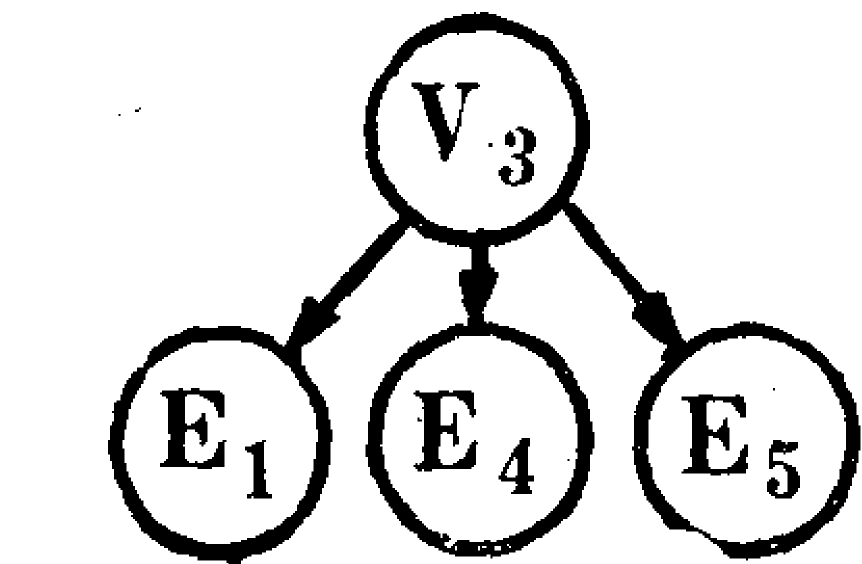
Der Lehrer gab dem Schüler ein Buch.

E_1 三价动词 E_3 E_4

（老师给学生一本书。）

Das Flugzeug verbindet die entlegen-

E_1 三价动词 E_4



sten Orte mit der Hauptstadt.

E_5

（飞机把偏远的地方同首都连接起来。）

由以上例句我们可以看出，虽然 Herringer 所划分的补足语只局限于体词性结构，但他特别强调了动词的放射作用。

另一位德国语言学家 Gehard Helbig 对动词配价问题作了进一步探讨。他把直接成分按照谓语动词对它们支配作用的强弱分为必要成分和非必要成分两种。认为只有必要成分才能成为配价结构中的补足语，而非必要成分都是自由说明语，它们都不是动补结构式中的补足语。例如上文引述的例句“Gestern besuchte mein Sohn mit einem Freunde das Museum”，句中的谓语 besuchte 为两价动词，Mein Sohn 及 das Museum 受谓语强支配，因而是必要成分，它们与谓语组成核心句；而 Gestern 与 mit einem Freunde 只受谓语的弱支配，因而是非必要成分，也叫自由说明语。另外，单就必要成分而言，补足语的配价结构还可分为必备配价和可选配价两种。请看以下例句：

Er lehrte 25 Jahre die Kinder

必要成分，必备配价 三价动词 非必要成分 必要成分，可选配价

die russische Sprache.

必要成分，必备配价

（他教了孩子们二十五年俄语。）

在动补结构式中，必备配价是不能省略的，省略了就要犯语法错误：

* Der Professor lehrt.

* Lehrte die Kinder.

然而可选配价是任选的，省略了不致造成语法错误：

Er lehrt die russische Sprache.

（他教俄语。）

这里需要强调的是，可选配价的补位仍是由动词的配价所分配的，它们属于必要成分，这和属于非必要成分的自由说明语是完全不同的。动词对自由说明语没有预示性的强支配作用，因而根本不存在省略问题。

Helbig 在他主编的动词配价词典中，分三个步骤为动词标注配价。第一步（Ⅰ）标明配价的数值，其中用括号括起来的是可选配价。第二步（Ⅱ）标明各种必要成分。各补足语之间用逗号隔开，同一补位的各变体形式用斜杠分开；其中不带括号的是必备配价，带括号的是可选配价。第三步（Ⅲ）进一步分别说明各补足语的语义和结构特征，并附有例句。

首先让我们以动词 Lehren 为例，看看这种动词配价词典的词条格式①。

①配价词典所应用的缩写词：
Sn＝第一格名词(Substantiv in Nominativ)，Sg＝第二格名词，Sd＝第三格名词，Sa＝第四格名词，Hum＝人物，Abstr＝抽象事物，Act＝行为，Anim＝生命体，I＝不带 zu 的不定式，Inf＝带 zu 的不定式，NS＝Nebensatz＝从句，NS daβ，W＝由 daβ或w- 疑问词 (wer, was, wen, wann, warum) 引导的从句。

Lehren

Ⅰ. lehren 2+(1)=3

Ⅱ. lehren→Sn, Sa_1/NS daβ, w/I/Inf, (Sa_2/Sd)

Ⅲ. Sn→1) Hum (Der Professor lehrt Sanskrit. 教授教梵文。)

2) Abstr. (Das Institut lehrt uns, die Zeit zu nutzen. 学院教导我们利用时间。)

Sa_1→1) Abstr (Er lehrt die Sprachen. 他教语言。)

2) Act (Er lehrt das Schwimmen. 他教游泳。)

NS→Act (Die Geschichte lehrt, daβ wir die Zusammenhänge verstehen/wer die Initiative zu exgreifen hat. 历史教给我们理解事物的内在联系/历史阐发谁应当作创始人。)

I→Act (Er lehrt ihn singen. 他教他唱歌。)

Inf→Act (die Geschichte lehrt uns, die Zusammenhänge zn verstehen 历史教给我们理解事物的内在联系。)

Sa_2→+Anim (Er lehrt die kinder, die Hunde das Reifenspringen. 他教孩子和狗跳环。)

Sd→+Anim (Er lehrt den kindern, den Hunden das Reifenspringen. 他教孩子和狗跳环。)

配价词典对动词 lehren 尚有如下的补充说明。

可选配价中的 Sa_2 与 Sd 可以互换：

Er lehrt die Kinder / den Kindern die Sprache. (他教孩子们语言。)

必备配价中的 Sa_1, NS, I, Inf 均可互换。值得注意的是，当I 占据必备配价位，且 Sn=Hum 时，Sa_2 则转为必备配价：

* Er lehrt singen.

Er lehrt ihn singen. (他教他唱歌。)

Not lehrt sparen, rechnen. （贫困使人节省和勤俭。）

下面再来看动词 übersefzen 的例子。其格式也分三步，但由于这个动词有两个义项，故分 a、b 两款分别列出。另外，这个动词的补足语还包括介词①。

übersetzen

a) übersetzen, übersetzte, hat übersetzt (=übertragen 翻译)

Ⅰ. übersetzen 1+(3)=4

Ⅱ. übersetzen → Sn, (Sa), (P_1 S), (P_2 S)

Ⅲ. Sn → Hum (Der Lektor übersetzt. 讲师在翻译。)

Sa → Abstr (Er übersetzt die Arbeit. 他翻译这部作品。)

P_1=aus

P_1Sd → Abstr (Er übersetzt aus dem Polnischne. 他译自波兰文。)

P_2 = in

P_2Sa → Abstr (Er übersetzt die Arbeit aus dem Deutschen in das Polnische. 他把这部作品从德文译为波兰文。)

b) übersetzen, setzte über, hat übergesetzt (=hinüberfahren 运载，摆渡)

Ⅰ. übersetzen 1+(2)=3

Ⅱ. übersetzen → Sn, (Sa), (pS)

Ⅲ. Sn → 1) Hum (Der Fährmann setzt über. 渡船工在摆渡。)

2) Hum（交通工具）(Die Fähre setzt über. 轮渡在运载。)

Sa → ±Anim (der Fährmann setzt die Wanderer, die

①p=介词，pS=由介词及其后面的名词组成的介词短语，pSd=支配第三格名词的介词短语，pSa=支配第四格名词的介词短语，Dir=Direction 方向。

Pferde, die Kisten über. 渡船工把游人、马和箱子摆渡过去。)

P = an , auf (方向性介词)

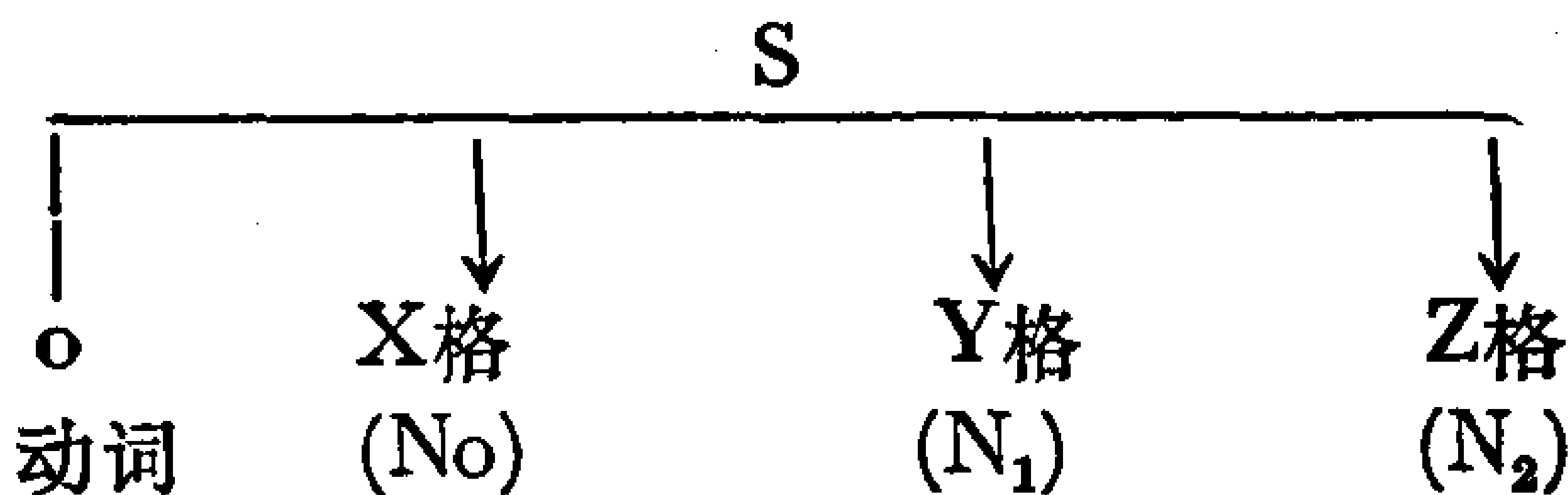
pS → Dir (Er setzt sie an das andere Ufer über. 他把她摆渡到彼岸。)

由以上例子我们可以看到，Helbig 在一定程度上解决了动补结构式的结构和语义特征的标记问题。他所提出的必要成分和非必要成分的划分，必备配价和可选配价的划分，对各种语言的动词研究都有参考价值。不过，由于他分析的对象是综合程度较强的的德语，所以在划分补足语时，不可避免地要借助于体词性结构的格范畴。

(2)深层结构格语法的配价系统

深层结构格语法是一种转换生成的理论，其中的一些观点对研究分析型语言动词的配价有一定帮助。从某种意义上说，深层结构语法是沟通综合型语言和分析型语言之间的桥梁。这种理论认为，核心句的基本结构是由一个动词和几个名词组构成的。各名词组都以某种格的关系与动词发生联系，而介词则是必要的格标记 (case marker)。

格语法所定义的核心句可以用如下的示意图表示：



以英语介词为例，常用的格标记至少包括以下几种：

主格 (Nominative): φ ; 施事格 (Agentive): by; 属格 (Genitive): of; 宾格 (Accusative): φ; 与格 (Dative): to; 裨益格 (Benefactive): for; 工具格 (Instrumentative): by, with; 处所格 (Locative): on, at; 在内格 (Inessive): in; 进入格 (Illative):into; 离格 (Ablative): from; 向格 (Allative): onto,

upon; 伴随格 (Comitative): with; 残缺格 (Abessive): without; 转变格 (Translative): into。

介词在英、法等分析型语言中十分活跃。在动补结构式中，动词往往支配一定的介词。因此，研究格语法对于探讨动补结构式的深层结构有一定的帮助。我们注意到，在充当格标记时，有的介词标示一种格，有的介词标示几种格，有的介词不标示格，而有些格却缺少介词作其标记，或者说只具备Φ型格标记。这种现象其实也是很自然的。在综合型语言中，不同的格具有相同的格词尾以及Φ型格词尾(秃尾)的现象不是也很常见吗？

深层结构格语法的配价系统认为，动补结构式包括变位动词及其三元补足语（也叫 participant, 参与者）。这三元补足语根据格的关系可分为高、中、低三个位。高位补足语 (N_0) 由主格表示，其人称和数受动词制约，因而作核心句的主语。中位补足语 (N_1) 与低位补足语 (N_2) 一般会具有与格、宾格或其他格的格标记。因而它们可以作核心句的间接宾语、直接宾语或状语。请看以下三个例句：

1) Votre jeune cousine donne à mon pauvre neveu un livre manifique.

   （你年轻的表妹给我可怜的侄子一本好书。）

2) Der Lehrer gab dem Schüler ein Buch.

   （老师给学生一本书。）

3) He gave the dog a bone.

   （他给狗一块骨头。）

从这些例子我们可以发现，这三种语言的间接宾语的结构形式尽管不同，但是它们都确确实实地以各自的方式占据着动补结构式中的一个格 (Dative) 或一个位(中位$=N_1$)。它们都是各自核心句中的参与者。

动补结构式中的必要成分常常包括某些受动词强支配的状语，它们尽管可以有多种结构形式，例如副词、名词、介词短语、

非限定动词、子句等等，但是从它们对动词的反作用的功能来看，它们都是修饰谓语的，因而在本质上有一种共性。按照格语法，我们可以把各类形式的状语都归纳为"状语格"①。

在英语的动补结构中，"前介状 A"是很常见的必要成分，这反映了分析型语言中的介词作为格标记所起的重要作用。例如：

The doctor concealed the fact from the public.

主A　　三价动词　宾A　　　前介状A

（医生对公众隐瞒了事实。）

按照英国语言学家 A.S. Hornby 的动词句型分类体系，《现代高级英汉双解辞典》为 conceal 这个动词标注了两种结构式②：Pattern 1 ＝动词＋直接宾语，Pattern 18＝动词＋直接宾语＋介词＋介词的宾语。由此可见，上面这个句子中的"主 A"和"宾A"都是必备配价，而"前介状 A"是可选配价。为同一动词标注两种以上的句式，实际上表明了动补结构式的兼容性及分布变体现象。需要强调指出的是这个句子中的"前介状 A"成分虽然是状语性的，但既然被纳入句型18。则说明它们受动词的强支配作用。从深层结构格语法的观点来看，作为格标记的介词 from 使名词组 the public 转换为"状语格"，具体地说，使这个名词组占据了一个离格的格位。

（3）P. le Goffic 制定的动补结构式

法国语言学家 Pierre le Goffic 等人采用类似于 Hornby 的方法对法语动词进行了动补结构式的分类研究。

首先，他根据后位补足语的结构状况制定动补结构式③。

---

①请对比 Lucien Tesnière 的从属关系语法理论。这种语法认为，动词的直接关联词无非是名词和副词，各种形式的状语均可转换成副词。

②A.S.Hornby 对英语动词作了较深入的研究。他将英语动词分为 25 个类型，并在《现代高级英汉双解词典》中为所有的动词标注了它们的类型号。他的这种分类体系实际上也是从配价的角度建立起来的。

③N＝名词；V＝动词；Inf＝不定式；Ind＝含直陈式动词的宾语从句；Subj＝含虚拟式动词的宾语从句；Adj＝形容词；直接结构＝不包含介词的结构；间接结构＝由介词引导的结构。

无后位补足语的动补结构式：

NV：Pierre marche.（Pierre 走路。）

带一个后位补足语的动补结构式，这个补足语可以是直接结构，也可以是间接结构：

NVN: Pierre aime Marie.（Pierre 爱 Marie。）

N V à N: Pierre sourit à Marie.（Pierre 向 Marie 微笑。）

N V de N: Pierre parle de Marie.（Pierre 谈到 Marie。）

N V Inf: Pierre sait conduire.（Pierre 会开车。）

N V à Inf: Pierre commence à travailler（Pierre 开始工作。）

N V de Inf: J'ai fini de travailler.（我停止了工作。）

N V que Ind: Pierre sait que Marie va venir.（Pierre 知道 Marie 要来。）

N V que Subj: Pierre veut que Marie vienne.（Pierre 希望 Marie 能来。）

带两个后位补足语的动补结构式，这两个补足语一个是直接结构，另一个是间接结构：

N V N à N: Pierre donne un livre à Marie.（Pierre 给 Marie 一本书。）

N V N de N: Pierre charge Marie d'un affaire.（Pierre 委托 Marie 一件事。）

N V N à Inf: Pierre oblige Marie à travailler.（Pierre 迫使 Marie 工作。）

N V N de Inf: Pierre excuse Marie d'être en retard.（Pierre 原谅 Marie 的迟到。）

N V à N que Ind: Pieerre promet à Marie qu'il viendra.（Pierre 答应 Marie 他来。）

N V à N que Subj: Pierre demande à Marie qu'elle vienne.（Pierre 请 Marie 要她来。）

带两个后位补足语的动补结构式，这两个补足语都是间接结

构：

N V à N de N: Pierre parle de Paul à Marie. (Pierre 向 Marie 谈及 Paul。)

N V à N à Inf: Pierre apprend à Marie à danser la valse. (Pierre 教 Marie 跳华尔兹舞。)

N V à N de Inf: Pierre demande à Marie de venir. (Pierre 请 Marie 来。)

带两个后位补足语的动补结构式，这两个补足语都是直接结构：

N V N Inf: Je regarde les enfants jouer. (我看到孩子们在玩。)

N V N que Ind: Pierre prévient Marie quil viendra. (Pierre 通知 Marie 他要来。)

带主语的表语的动补结构式：

N V Adj: Pierre paraît contont. (Pierre 显得很高兴。)

N V N : Pierre est devenu mécanicien (Pierre 成为机械师。)

带宾语的表语的动补结构式：

N V N Adj: Pierre rend Marie heureuse. (Pierre 使 Marie 很高兴。)

N V N N : Je m'appelle Pierre (我叫 Pierre。)

带有主宾共动式的对称动词的动补结构式：

V $N_1=N_1$V Je colle un timbre — Le timbre colle. (我贴邮票——邮票贴上了。)

鉴于谓语动词向它的两侧都有放射功能，P. le Goffic 对动补结构式中的名词，其中也包括前位补足语——主语，都作了某些语义及结构上的说明。例如动词 oser (敢)、espérer (希望)要求其主语是人 (qqn)；动词 rouiller (生锈)、étinceler (闪耀)要求其主语是物 (qch)；而动词 apporter (带来)则要求其直接宾语是物，间接宾语是人或是物。另外，动词的配价还体现在有

的动词要求其主语采取无人称形式。在法语中主位无人称结构还包括外位式与分离式。

外位式 (extraposition)，用 il 充当形式主语，真正的主语后置，例如：

Il passe souvant des eamions dans cette rue. （这条街上经常走卡车。）

分离式 (détachement)，用 ça 充当形式主语，真正的主语置于逗号之后：

Ça m'a coûté 500F d'aller en vacances.（去度假花了我 500 法郎。）

P. le Goffic 按动补结构式及名词语义特征列出了许多表格，在每个表中列出了符合该表结构式的常用动词，并举例加以说明。下面仅以表 15 为例作一个简单介绍。

表15　基本句型：Pierre donne un livre à Marie.

后位补足语 / 动词	ϕ	N		N à N	其他结构
		qch	qqn	qch à qqn	
apporter (带来)		+		+	qch à qch
approndre II(数)				+	à N à Inf.
apprendre IV(告诉)				+	à N qne Ind
chanter (唱)	+	+		+	
coûter (花费)		+		+	主语可用分离式
dessiner (画)	+	+	+	+	

其中，coûter 一词有如下的用法，举例说明：

N V N: Ce manteau coûte 500F.

（这件大衣价值 500 法郎。）

N V N à N: Ce manteau a coûté 500F à Marie.

（这件大衣花了 Marie 500 法郎。）

分离式：Ça m'a coûté 40F, de mettre ma voiture là.

（把车子停在那里花了我 40 法郎。）

六五三四 出二

我们看到，由于 P. le Goffic 所制订的动补结构式包含一些初步的逻辑语义信息，因而在一定程度上揭示了动补结构式的深层意义。但是，诸如上列动词 apporter 和 coûter 所支配的直接结构补足语在功能上的差异，只有采用更多的逻辑语义信息才能分清。

(4) Maurice Gross 对动词配价的矩阵表分析

法国另一位语言学家 Maurice Gross 对法语动词的配价与动补结构式研究得更为深透。他采用矩阵表的形式分析了动词各式各样的配价特征。这种做法的特点在于很好地体现了动词配价结构的兼容性以及各种结构式之间的内在联系，合理地安排了介词的格标记问题。他将 3000 个法语动词以义项为词条单位分列在 19 张矩阵表中。每一种矩阵表代表一两种基本的动补结构式，但由于矩阵表上的配价兼容性处理得好，实际上每个表可以反映好多种动补结构式。由于他用 100 个左右配价特征来分析动词，所以分类极为细微，3000 个动词竟然分成了 2000 个细类，每一类平均只有 1.5 个动词，也就是说，没有两个动词具有完全相同的配价结构。

现将 Gross 的 19 个矩阵表所处理的基本结构式及各表所包括的配价特征位列表总结如下(见表 16)[1]：

表 16

矩阵表编号	主要动补结构式	配价特征位
1	N_0 U Prép $V^\circ\Omega$	25
2	N_0 V $V^\circ\Omega$	28
3	N_0 V N_1 $V'\Omega$	28
4	QuP V N_1	19
5	QuP V Prép N_1 以及	27

[1]N 表示体态动词，Qup 表示从句，V°与 V' 均表示不定式，右上角的数字表示其逻辑主语是 N_0 还是 N_1，Ω一般表示不定式的连带成分。

	I_t V Prép N_1 QuP	
6	N_0 V QuP	36
7	N_0 V à ce QuP	29
8	N_0 V de ce QuP	28
9	N_0 V QuP à N_2	45
10	N_0 V QuP Prép N_2	33
11	N_0 V N_1 à ce QuP	34
12	N_0 V QuP 以及 No V N_1de V'Ω	30
13	N_0 V N_1 de ce QuP	33
14	N_0 V à ce QuP Prép N_2	33
15	N_0 V de ce QuP Prép N_2	42
16	N_0 V $Prép_1$ N_1 $Prép_2$ N_2	42
17	I_1 V Prép ce QuP Prép N_2	24
18	N_0 V Prép N_1 Prép N_2 Prep QuP	39
19	Qup V N_1 Prép N_2	19

由于表格很多，下面仅以表 17 为例作一个简要的介绍。首先谈谈这个表里的 42 个配价特征位的涵意。

第 1 位到第 5 位是描述主语 N_0 的：其中第 1 位要求主语为表示人的名词，第 2 位表示对作主语的名词不加语义限制，第 3 位是要求主语为带限定句的形式，第 4 、5 两位是要求以不定式作主语，第 6 位是以 N_0 作主语的被动结构，第 7 位是动词不及物的用法。

从第 8 位到第 24 位给出直接补语或间接补语 (Prép) N_1 的各种配价结构①。究竟是直接的还是间接的，要看第 9 位上是 $_0$ 还是给出具体的介词。第 8 位结构式中的 Prép 究竟有无也要参看第 9 位。从第 10 位到第 18 位给出 N_1=子句的情况，其中第

①Gross 在此处所定义的补语表示后位补足语，其中不带介词的为直接补语，带介词的为间接补语。很明显，这同中介成分中所定义的直接成分、间接成分具有完全不同的概念。

10 位表示直陈式从句。第 11 位表示虚拟式从句。第10 位与第11 位之间的分隔线没顶到头。表示这两位的情况还涉及第 12 位至第 16 位的几种结构变体。其中第 12 位的 [pc.z] 表示可以省去引导补语从句的介词和 ce①，第 13 位表示带 si ... ou non的子句②。第 12 位上方向第 13 位拐进，表明第 12 位的特征也施加于第 13 位，因此第 13 位所表示的子句也不用介词和 ce 引导。

第 14 位至第 16 位都是指由子句缩略成的不定式结构作中位补足语 (N_1)，它们之间的区别仅在于逻辑主语不同以及是否由介词引导。第 17 、18 位表示从句或不定式的代词化，第 17 位包括 ceci 或 cela，第 18 位表示置于谓语动词之前的各种宾格或"状语格"的代词③。

从第 19 位到第 23 位表示中位补足语 (N_1) 为名词、代词、带限定从句的结构。第 19 、20 位表示人物名词及其代词化，第 21 位要求并非指人的名词，第 22 位及 23 位均属于第 21 位的变体，第 24 位是外位被动转换规则④。

从第 25 位到第 42 位都是描述间接补语 Prép N_2 的。由于第 26 位上都有具体的介词，所以第 25 位上的介词 Prép 也总是非空缺的。由于这一区间的问题都同第 8 位到第 24 位相似，故不再赘述。只是第35 位上规定了以 là 作代词的形式，第 42 位上给出低位补足语前置的结构。

---

①Paul est content de ce que Marie vienne. [pc.z] Paul est content que Marie vienne. (Marie 能来，Paul 很高兴。)

②Paul saura si Marie est arrivée ou non. (Paul 知道 Marie 是否到来。)

③ppv = pronom préverbal，根据具体情况尚可分为 ppv le (la, les), ppv lui (leur), ppv en, ppv y 等等。

④规则施加对象: No V QuP (1) On a affirmé que Marie était ici. (我们肯定 Marie 在此。)

被动转换 [passif] ⟶ Qup est Vpp par No (2) Que Marie était ici a été affirmé par Luc. (Luc 肯定 Marie 在此。)

外位转换 [extrap] ⟶ il est Vpp par No QuP (3) Il a été affirmé par Luc que Marie était ioi. (Luc 肯定 Marie 在此。)

表 17

Sujet 1 Nhum	2 Nnr	3 le fait Qup	4 $V^1\Omega$	5 $V^2\Omega$		6 N_0 est $Vpp\Omega$	7 N_0 V	8 N_0 V Prép N_1	Comlèments directs ou indirects 9	Complétives 10 que P	11 que P Subj	12 [pc z.]	13 si P ou Si P	14 $V^\circ\Omega$	15 $V^2\Omega$	16 de $V^\circ\Omega$	Pron 17 ce (ci + la)	18 ppv
+	–	–	–	–	abstaire	–	–	–	0	+	–	–	–	+	–	–	+	+
+	+	+	–	–	ajouter	–	–	–	0	+	+	–	–	–	–	–	+	+
+	–	–	–	–	alterner	–	–	–	0	–	+	–	–	–	–	+	+	+
+	–	–	–	–	s'apercevoir	–	–	+	de	+	–	+	+	+	–	–	+	+
+	+	+	+	+	apprendre	–	–	+	à	–	–	–	–	–	–	–	–	–
+	–	–	–	–	compter	–	–	+	sur	–	–	–	–	–	–	–	–	–
+	+	+	–	–	définir	–	–	+	0	–	–	–	–	–	–	–	–	–
–	+	+	–	–	différer	–	+	+	de	–	–	–	–	–	–	–	–	–
+	–	–	–	-	embaucher	–	+	+	0	–	–	–	–	–	–	–	–	–
+	+	+	+	+	enseigner	–	+	+	a	–	–	–	–	–	–	–	–	–
+	+	+	+	+	extraire	–	–	+	0	+	+	–	+	+	–	+	+	+
+	+	+	+	+	identifier	+	+	–	0	–	+	–	–	–	–	+	+	+
+	+	+	+	+	montrer	–	–	–	a	–	–	–	–	–	–	–	–	–
+	–	–	–	–	préferer	–	–	+	0	–	+	–	–	+	–	+	+	–
+	–	–	–	–	se rendre compte	–	+	+	de	+	–	+	+	+	–	–	+	–
+	+	+	+	+	sortir	–	–	–	0	+	–	–	+	+	–	–	+	–
+	–	–	–	–	tirer	–	–	–	0	+	–	–	+	+	–	–	+	–
+	+	+	+	+	tromper	–	+	+	0	–	–	–	–	–	–	–	–	–

表 17（续）

							Compléments indirects																
Noms					24	25	26	Complétives									Noms					41	42
														Pron									
19 Nhum	20 ppv	21 N-hum	22 le fait QnP	23 ppv	24 [extrap] [dassif]	25 N. V prép N_2	26	27 que P	28 que P subj	29 [pc z]	30 si P ou Si P	31 V°Ω	32 V'Ω	33 ce (ci+la)	34 ppv	35 là	36 Nhum	37 ppv	38 N-hum	39 le fait Qup	40 ppv	41 [extrap] [pasif]	42 prép N_2N. V N_1
−	−	+	−	+	−	−	de	+	−	−	−	−	−	+	+	−	−	−	+	+	+	−	+
−	−	+	−	+	−	+	a	−	+	−	−	+	−	+	+	−	+	+	+	+	+	+	−
+	+	+	−	+	−	−	avec	−	+	−	−	+	−	+	−	−	+	−	+	−	−	−	−
−	−	+	+	+	−	−	a	+	−	−	−	−	−	+	−	+	−	−	+	+	−	−	−
+	+	−	−	−	+	+	a	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−
+	−	+	+	−	−	−	pour	−	+	−	−	−	+	+	−	−	−	−	+	−	−	−	−
−	−	+	−	+	+	−	pour	−	+	−	−	−	+	+	−	−	−	−	+	−	−	−	−
+	−	+	+	+	−	+	en	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−
+	+	−	−	−	−	+	pour	−	+	−	−	−	+	+	−	−	+	−	+	−	−	−	−
+	+	−	−	−	−	−	a	−	−	−	−	−	+	−	−	−	−	−	−	−	−	+	−
+	+	+	−	+	−	−	de	+	−	−	−	−	−	+	+	+	+	+	+	+	+	−	+
+	+	+	+	+	+	−	a	−	+	−	−	+	−	+	−	−	+	−	+	+	−	−	−
+	+	−	−	−	−	−	a	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−
+	+	+	+	+	−	−	a	−	+	−	−	+	−	+	−	−	+	+	+	+	−	+	−
−	−	+	+	+	−	−	a	+	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−
+	+	+	−	+	−	−	de	+	−	−	−	−	−	+	+	+	+	−	+	−	+	−	+
−	−	+	−	+	−	−	de	+	−	−	−	−	−	+	+	−	+	−	+	−	+	+	+
+	+	−	−	−	−	+	en	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−

下面我们再来看一看这类矩阵表的使用方法。总的说来，凡是某一个动词具有某一位的配价结构时，则在其交叉处标一个“十”号，否则的话，则标一个“-”号。这张表共分析了三元补足语，也就是高位（主位 N_0）、中位（N_1）与低位（N_2）补足语，对于任何一个动词来说，如果我们要了解它有怎样的必备配价和可选配价，就必须注意第7位、第8位及第25位所提供的信息。如果某动词在这三位上都是“-”号，就说明省略后位补足语或单独出现中位或低位补足语都是不允许的，也就是说，在这个动词的动补结构式中，N_1 与 N_2 都是必备配价，它们中的任何一个都是不能省略的。反之，如果某动词在这三位上都是“+”号，则表明 N_1 与 N_2 都是可选配价。如果一个动词在第7位上是“-”号，在第8位上是“+”号，在第25位上是“-”号，则表明 N_1 是必备配价，N_2 是可选配价。如果一个动词在第7、8两位上都是“-”号，在第25位上是“+”号，则表明 N_1 是可选配价，N_2 是必备配价。另外，需要补充的是，如果一个动词在第7位上是“+”号，在第8位及第25位上都是“-”号，则表明 N_1 与 N_2 是共轭必备配价（或则一起出现，或则都不出现）。如果一动词在第7位上是“-”号，在第8、第25位上都是“+”号，则表明 N_1 与 N_2 是互补必备配价（二者至少必备其一）。

由于这种矩阵表对配价的必备和可选特征及对介词性格标记处理得很巧妙，因而结构式的分布变体有很大的兼容性。它所能提供的动补结构式远远超出了其基本结构式的范围。但是，结构的兼容性是以义项为基础的，一张表内的一个动词一般只加工它的一种义项结构，如果这个动词尚有其他义项，则需列在别的表内另行加工。例如动词 apprendre 包含有四个义项（Ⅰ学，Ⅱ教，Ⅲ听说，Ⅳ告诉），因而放到四张表中分别加工。

我们可以看出，表16所加工的 apprendre 是它的第2个义项结构。第32位上给出 $V'\Omega=$“+”，表示不定式的逻辑主语是 N_1，例如：

Pierre apprend à Marie à danser la valse.

N_0 V N_1 V' Ω

(Pierre 教 Marie 跳华尔兹舞。)

显然，这和 apprendre I 的义项配价结构(在表 7 中另行加工)是不同的：

Pierre apprend à danser la valse.

N_0 V V° Ω

(Pierre 学跳华尔兹舞。)

下面让我们分析一下动词 Préférer 的结构。由于这个动词在第 7 位上是“—”，在第 8 位上是“+”，第 9 位上是“0”，第 25 位上是“—”，于是可以得知 préférer 的直接补语 (N_1) 是必备配价，而间接补语 (N_2) 是可选配价。另外，由于这个动词的直接补语既可以是名词性的，又可以是从句性的，所以我们能够生成下面这样的核心句：

Je préfère le bleu au vert. (我喜欢蓝甚于绿。)

Je préfère qn'il parte. (我宁可他走。)

最后让我们分析动词 embaucher 的间接补语的结构。通过查表可以看到，这个动词第 28 位及第 32 位上的“+”号表明它的后位补足语 N_2 既可由虚拟式从句充当，又可由不定式充当：

Pierre embauche Paul pour qu'il fasse ce traveail.

N_0 V_2 N_1 Psubj

Pierre embauche Paul pour faire ce travail.

N_0 V N_1 V' Ω

(Pierre 聘请 Paul 做这项工作。)

通过这个例子可以看出，矩阵表对兼容结构式的成分和关系处理得很好。上面句子中的不定式结构是作为子句的结构变体或缩减式来处理的，因为 N_1V' 已构成了类似于子句的主谓结构，只不过 V' 是非限定动词罢了。

以上介绍了 M. Gross 对动词配价的矩阵表分析。近年来，他的工作又有了新的发展。我们认为，在研究动词的配价方面，他的系统更为精细、严密和形式化，因而对自然语言的自动处理有较大的参考价值。

（5）动词配价论和谓语中心论的异同

在分析探讨了国外的几种动词配价论之后，我们认为，这种理论的精华就在于强调了动词的支配作用、放射作用和轴心作用。从这种意义上说，动词配价论同我国机器翻译界一贯倡导的谓语中心论确有一些共同之处，不过，后者采取了一种十分独特的分析体系，因而似乎更能揭示语言结构的层次和动态转换机制。动词配价论与谓语中心论的共同之处在于都强调谓语动词对其他成分的支配、主宰作用。不同之处在于，动词配价论一般只对原语直接成分作独立分析，动补结构式也集中于探讨必要成分的形式和数量；而谓语中心论是对原语的各级层次作相关分析，不仅要探讨动补结构式中的必要成分，还要探讨各种类型的自由说明语，因为从翻译的角度来说，无论是强支配成分还是弱支配成分在施加转换方面的可能性和必要性都是平等的。

我们之所以对动词配价论感兴趣，就在于这种理论研究了我们所关心的问题的一个重要方面，它能够向我们提供动补结构式的静态规则。我们的任务不仅在于了解动词的配价和动补结构式，还在于为这些结构式制定转换式。如果我们把所有这些参数都存贮到机器词典中予以预制的话，那么它们不但可以为相关分析提供更多的预示信息，还可以为转换和综合提供更加便利的条件。

5.4 机器翻译中的一些语言学问题 在机器翻译中，除必须加强动词的研究之外，还有一些其他的语言学问题必须认真加以对待。这主要包括：词的分类问题，连词和标点的处理，以及介词的研究。下面分别进行讨论。

5.4.1 词的分类问题 词是自然语言中的基本单元，同时也是自然语言处理中的基本单元。词的分类信息是计算机加工的初始信

息。分类是否合理直接影响到计算机工作的成效。因此，词的分类问题是机器翻译和其他自然语言处理研究中的重大问题。

根据目的和要求的不同，词的分类可以各种各样。一般来说，根据语法特征分类和根据语义特征分类，是各类信息处理所共同需要的。针对机器翻译的特点，词的转换信息必须配置。对于人机对话来说，“知识信息”或“非语言信息”又是必不可少的。

传统语言学对词进行的语法分类和语义分类，都是含混的，缺乏形式化的标志，不能满足机器翻译等自然语言处理研究课题的需要。因此。我们从研究机器翻译的第一天起，便对词进行了独特的分类。随着研究的深入，词的分类问题不断改进。1961 年的俄汉机器翻译方案中已赋予三种功能不同的组别，即语法组别、语义组别和对应组别。多年来，我们发展了一套“类属组”的词的分类法。有的方案为了更充分地照顾语义分析的需要，还设立了“语义场”。

看来，对于机器翻译等一类语言工程来说，比较合理的分类法是扩大分类信息的容量，并把它们划分为三大类信息：语法信息、语义信息和转换信息。每一类信息在必要时还可分为三级类属组。语法信息包括形态、结构(分布和支配关系)和功能（即句法成分)三小类。一、二小类属静态特征，即词的固有信息，其中第一小类指本身的形态特点，第二小类指支配它词或受它词支配以及其他分布特点，第三小类属动态特征，即根据上下文而定的信息。语义信息包括人或动物名、时间名、空间名、抽象名、专名等等(以上属名词)；智力、替代、形成、相关、趋向、连动、赋予、感受、意愿等等(以上属动词)。转换信息包括移位、换位、加减词和修辞等等。这样做有以下几点好处：(1)满足了增添信息的需要，(2)求得了内部一致性，(3)照顾了与其他自然语言处理的联系。比如说，只研究英语语法的人基本上可以在这种语法信息和语义信息的基础上进行工作，同样地，用计算机研究自然语言的人划分出来的语法信息和语义信息也可以为机器翻译服

务。

在进行词的分类时，应该强调一下经济和足够的原则。分得太少，不便于分析工作。分得太多，会产生矛盾交叉现象，给分析工作造成麻烦。看来，系统的扩大将增加按词项分类的必要性。

目前我们把词分为 14 大类：名词 (N)、形容词 (A)、代词 (R)、介词 (P)、限定词 (D)、副词 (F)、助动词 (X)、实义动词 (V)、现在分词 (G)、过去分词 (E)、不定式 (I)、同等连词 (C)、非同等连词 (L)、小品词 (K)。

用以表示词的配价结构的分布式可视为对词类的一种补充。例如除主语之外，尚支配两元的动词可能有如下的分布式（此处 S＝从句，Y＝不带小品词的不定式）：

VNN, VNI, VNY, VNS, VNG, VNE, VNF, VNA, VNP, VPP, VPI,VPY, VPS.

支配参数和语义场参数又可形成对分布式的进一步的说明。例如动词 deprive 的分布式是VNP，其中N的语义场是 HUMAN，P 的支配参数是介词 of。利用分布式作为词的分类补充，很容易记载向译语的转换信息。又如根据动词 conceal 的分布式 VNP (P 为介词 from) 可构成其转换式：“对 PVN”。这里，“对”为介词词义。

语义场参数还可与语义分布式配合使用，从而实现判别多义的功效。例如英语动词 receive 的两个义项（“接待”、“收到”）在词典中占有两个记录。它们的分布式都是 VN，但语义场参数不同。动词的义项为“接待”时，语义场中有表示人物特征的语义因子；动词的义项为“收到”时，语义场中则有表示物体特征的语义因子。

5·4·2 连词和标点的处理　语符流的进一步切分是句子的切分，切分不对，就谈不到正确的理解和翻译。对于书面语来讲，句子的切分主要靠连词和标点。如果是口语，标点就要让位给停顿、语调等语言特征。另外，还有一些关联词语，如“一边……一边”、

“不是……而是……”，也是切分句子的好标记。

关于标点的处理，这里先着重谈一下词的处理问题（以英语 and 为例）。前些年，我们对 and 进行过一番研究。据粗略统计，and 在五千多条题录中出现1300多次（即平均每四句题录中就有一个 and）。又据前两千句统计，and 共出现676次（即平均每三句题录中有一个 and），其中作非同等连词和准同等连词使用的有57个（占 and 总数8.4%），作同等连词使用的有619个（占 and 总数的91.6%）。

以上数字表明，如果 and 一律处理为同等连词，则676个句子中有57句是错误的。如果对 and 进行处理，但未处理好，则错句的量数可能小于或等于57句，还可能大于57句。由此可见分析 and 的重要性。

and 问题的复杂性主要在于它的三种功能（同连、非同连、准同连）不好分辨。对三者的分辨完全靠语法手段行不通，必须辅以语义手段，而语义手段有时也不那么可靠。通过摸索，我们认为，采取分散方式解决 and 比较好，这就是：

（1）有些词组在成语词典和分离结构词典中解决，如 now and then, relation between ... and;

（2）联结形容词、副词、动词的 and 在加工相应的词类或 and 的第一次加工时解决，如 adj and adj, adv and adv, ving and ving, ved and ved, vinf and vinf;

（3）与逗号连用的 and，在 and 第一次加工时解决，如 NP, NP and NP;

（4）联结介词的 and，在 and 第二次加工时解决，如 NP and of, by and from, of and NP;

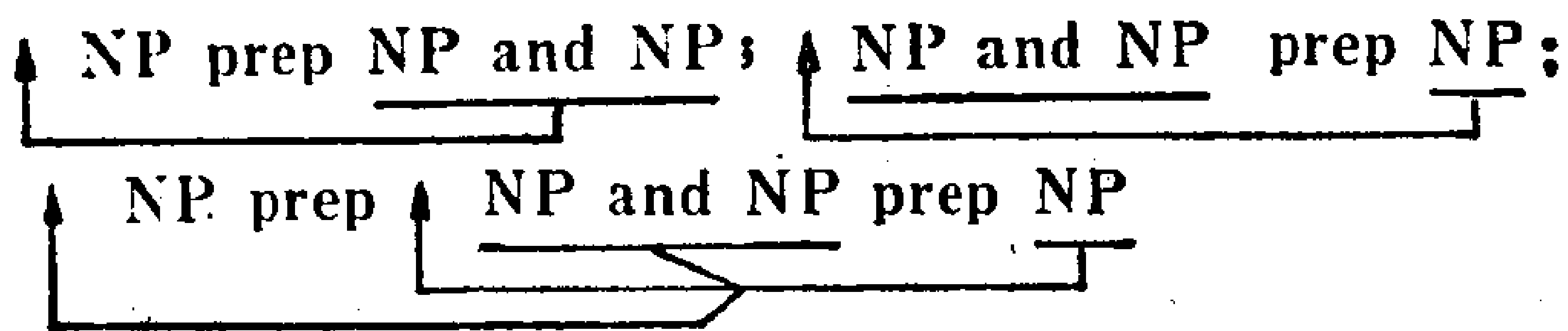
（5）联结名词的 and，在第二次加工时解决，如 NP and NP。

上述几类 and 当中，NP and NP 是 and 问题的核心。对于它，也以采取分段解决为宜。先查清这个词组周围是否有介

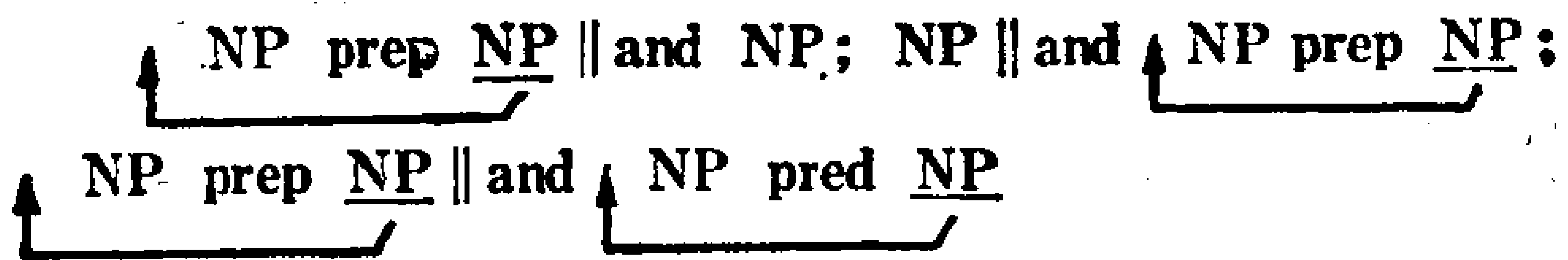
词，介词在前还是在后或前后都有介词，然后分段加以解决。另外，在进行这种处理之前，先查清前面 NP 或后面 NP 的一些带规律性的条件，也很必要，因为这可以不必再找更大范围的条件便可给出正确结论。这里可利用前面 NP 中的冠词，还可利用后面 NP 中的 the、its、their、other 等。

通过分析，可给 and 得出下列三种结论：

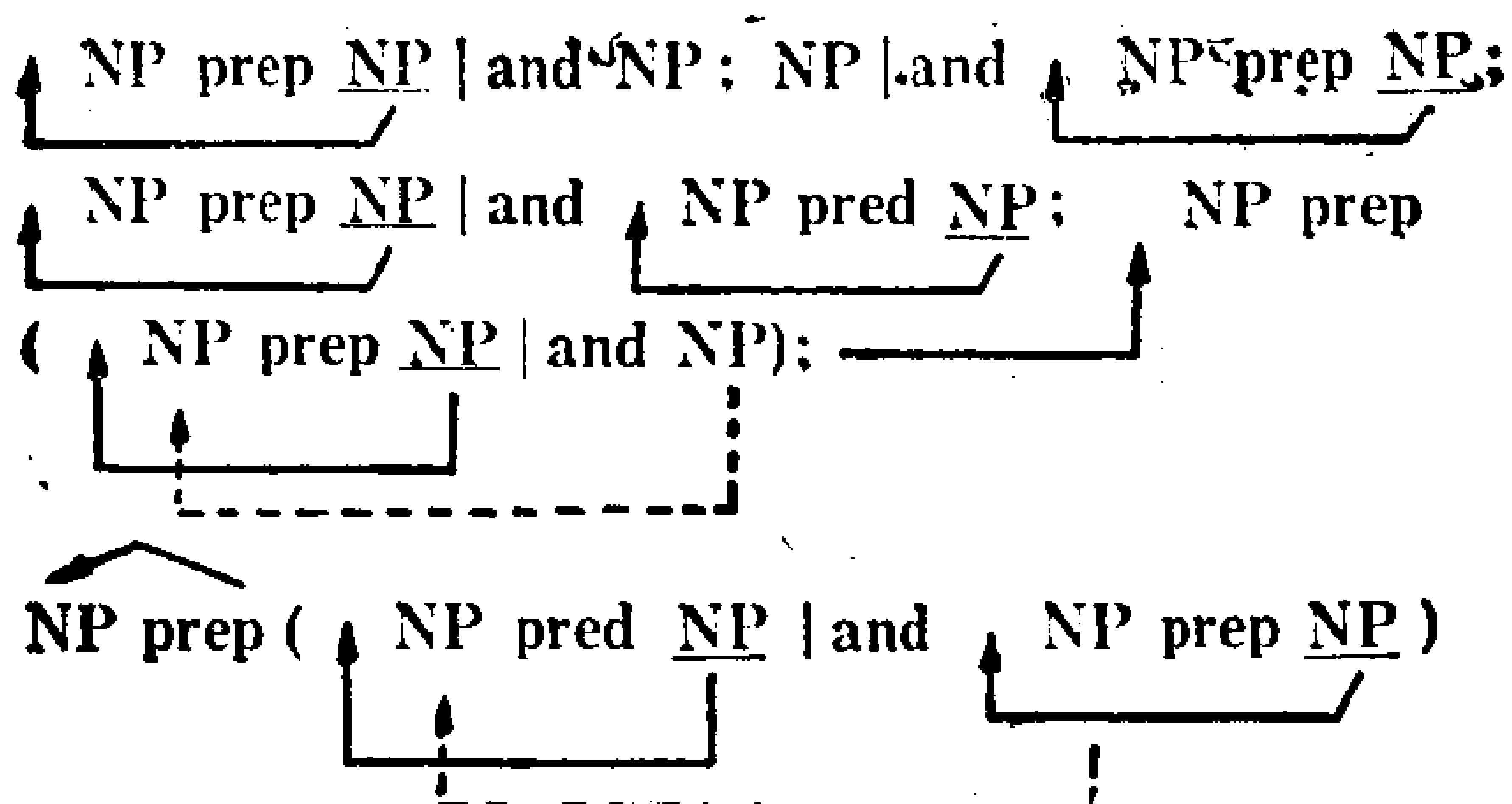
(1)同等连词。这种连词所联结的前后两个 NP 可以连在一起，给同号，其类型为：NP and NP;



(2)非同等连词。这种连词所联结的是两个相对独立的手段，而不是它前后的两个NP。在全文翻译中，它一般联结两个分句，而在题录翻译中，则联结两个句段。因此，可以把它看作把前后截然分开的标记，不能给前后两个NP同号。其类型为：



(3)准同等连词。这种连词介乎上述二者之间，它所联结的既不是前后两个 NP，也不是相对独立的句段。它联结的是有一定关系的短语，而这种一定关系常常表现在一部分“准同连”，要找到它所联系的词(为此设立“带准同连”)，以便在翻译移位时把两个短语连起来(即给同号)同时移。其类型为：



前面已经谈到，一旦语法手段不能分辨出这三种不同功能时，要辅以语义手段。所谓语义手段，就是比较前后 NP 的中心词或前 NP 的中心词与后 NP 中的某个 N 是否在语义类别上一致。一致时判为“同等成分”，and 得“同连”特征；不一致时另作别论。但是在“查对同类”方面还有许多可以探讨的问题。拿一个普通的例子来说，N and NNN，一般的理解可能有 N and NNN 和 N and NNN 这两种同等。但在英语中有时也常会碰到 N and NNN，N and NNNN，甚至 N and NNNN 的情况，这是受英语的一种省略法制约的。指出这一点，对正确理解原文和查对同类很有帮助。

与之同时应该指出，根据“查对同类”判别同等成分的方法还很不完善，因为在题录中还可能有以下情况，即看起来截然无关的两个 NP 应该看作同等成分连起来，相反地，看来有关的两个 NP 又不能作为同等成分连起来，甚至是可连可不连。

(1) production and properties of a …

(2) … principles || and applications to …

(3) in steel and specialty steel production（这里两个 steel 同等。）

(4) man || and machine in modern society（这里可连也可不连，有二义性。）

从上面所述可以看出，小小的 and 竟然是计算机理解自然语言研究中的大问题，尤其是对我们这种词序差别较大的语言来说。试比较：

（俄）Надо изучать историю и грамматику китайского языка.

（英）It is necessary to study history and grammar of Chinese language.

（法）Il faut étudier l'histoire et la grammaire de Chinoise.

（汉）必须学习历史和汉语语法（或"必须学习汉语历史和语法"）。

有人曾建议来一个"模糊译法"，即先译后面再加前面，得"汉语语法和历史"译文，任你自己去理解。这恐怕也不是一个好办法。

连词问题的确不是个小问题，不能等闲视之。这里谈的，只是一些思路，离问题的解决还有不小的距离，说不定，经过很大努力，还不能得到百分之百的解决呢。这一问题我们在机器翻译研究的初期便已提出，到今天仍未能比较彻底地解决，就是一个证明。

5·4·3 介词研究　介词是一种格标记，它与动词的配价有着十分密切的联系。虽然介词属于一种虚词，但它的作用并不亚于实词。实词和虚词的划分，主要是根据能否表达实体意义。一般说来，人们对于实体意义的理解并不困难，而介词的结构意义、关系意义纵横交错，却很不易掌握。介词不仅是指示句中实词之间有某种关系的词，而且借助本身的意义，进一步揭示和确立这些关系的内容和性质。如果没有介词或介词的等价物（形态丰富的语言用格的形式来代替某些介词，日语是用助词来代替），人们也就只能表达一些简单的概念。介词在句子中起着重要的媒介作用，许许多多的介词又组成各种各样的短语，它们的句法作用不尽相

同，它们的意义又变化多样（at、by、in 等常用介词在牛津词典中的注释平均有 365 个义项）。因此，介词的处理，对于人们正确理解自然语言是个重要问题，对于机器理解自然语言，更是一个关键。可以毫不夸张地说，目前机器翻译中出现的错误，多数是由于介词处理不当而造成的。

从机器翻译的要求来说，处理介词必须要明确以下五点：

a) 介词在源语中所处的位置；

b) 介词在译语中的位置；

c) 介词分布的层次划分；

d) 介词的句法功能；

e) 介词的词义。

机器加工介词，是一个十分复杂的问题，需要一定的条件。比如，它要求在加工介词之前，谓语和偏谓语都已确定，各词词组已经形成等。另外，介词的及早确定，对于分析其他词也有一定的好处。根据这种情况，让机器采取分段加工介词的办法，看来是颇有成效的。

(1)成语处理

所谓成语处理，就是针对介词短语数量多的特点，首先将已经比较固定的词组选出，定作成语。通过机器的成语词典首先将介词进行成语处理，不仅迅速简便，而且可以为进一步分析综合准备一部分参考信息。

这类成语的意义是单一的，语法作用也容易确定（有的本身就已明确，有的尚需一些简单的条件来确定），而且完全可以作为一个单词处理。考虑到机器在翻译时的方便，我们可以尽量将一些经常使用的成语词组收集起来，供机器使用。例如：in turn（依次）[状语]，in order to（为了）[等于引起状语的 to]，consist of（由……组成）[一般用作谓语]。

有些成语中加入了其他的词，但是有其明确的词义和句法功能（例如，下面的成语都用作状语）。为便于机器查找，干脆一一

收入，对号入座。例如：

in different way （以不同方式）

in the same way （同样）

in the following way （按下述方式）

in very nearly the same （以极其相似的方式）

in one way or another （用种种方法）

in the way explained below （以下要谈到的方法）

（2）结构处理

语言中常有一些由介词与其他词组成的固定结构格式。这种结构格式在性质上与成语相同，但在形式上与成语不同，它们中间可以嵌套其他词，因此可称之为“分离结构”。充分利用这种类型的搭配关系，对于解决介词问题十分有利。例如：

effect of … on … （……对……的影响）

effect =偏谓，“影响”

of =前介主 B，“零义”

on =前介宾 C，“对……的”

（3）图示处理

经过成语和结构处理后，剩下来的就是一般的介词短语了。无论从数量上还是从问题的复杂性上讲，这一部分的工作量都远远超过前两部分。因此，编好介词图示是解决介词问题的中心任务。

由于介词具有功能多样和语义丰富的特点，因此介词的分析极其复杂。分析介词需要多种条件，一般至少要有另外两个或三个条件。例如“in”，单了解其后是什么词是无法得到正确译文的。只有把两者结合起来，才能得到正确的译文。

当然，上面谈的仅仅是一般情况，遇到特殊的情况，还需要更多的条件。现在让我们来看看图示处理中对介词的具体分析和加工。

第一，查问介词是否位于句首。介词短语有作定语、状语和宾语的可能，但如果该短语出现在句首，就排除了作定语的可能

性，绝大多数情况下是作状语，少数可能是倒置宾语。进一步查问与谓语有无搭配关系，就可能定其是否为状语。例如：In 1905 Albert Einstein formulated a theory about ...（在 1905 年，爱因斯坦陈述了一种……理论）。

第二，查问介词是否与动词、动名词、形容词及个别副词有搭配关系。在 devote (oneself) to, suitable for applicable to, independently of 等结构中，介词 to、for 和 of 的分析，只需找到其支配词就可以确定。例如：There are two types of heat treatment applicable to aluminium alloys.（有两种适合于铝合金的热处理方法。）

第三，查问介词前面有无名词。如果介词前没有名词，一般就不会是作定语，只可能作状语或定语。但是要注意，英语中有所谓"分隔型"宾语，这种定语在科技文献中是比较常见的。其特点有三：一是它在谓语（通常是被动语态）的后面，它的被定语却在谓语前面；二是被定语和介词之间往往有某种搭配关系；三是定语一般较长，宜置于句子后面。例如：Details are given of a steel works having a design capacity of 350,000 tonne / month,（给出了具有设计能力为 350,000 吨/月的钢厂的细节。）

第四，查问介词本身的特征。有的介词有多种功能，而有的只有一种或偏重于一种。有的介词意义很多，而有的基本上是单义的。根据这些特点，再加上具体情况的分析，就不难得出应有的结论。例如 during 在绝大多数的情况下，充当状语，但在个别场合也可以作定语：

The light during the day is given by the sun.

（白天的光亮是太阳给予的。）

第五，查问介词是在谓语前还是在谓语后。介词短语前为名词，又都处于谓语前，一般为定语。处于谓语后的介词短语，如果其前面没有名词，应为状语（"分隔型"定语除外）。如有，则需根据其他条件进一步分析，辨别它是名词的定语，还是谓语的状

语。如下面两个例子：

The bell at eleven o'clock tells us that the class is over.

（11点的钟声告诉我们下课了。）

They ring the bell at eleven o'clock to tell us that the class is over.

（他们11点时打钟告诉我们下课了。）

第六，查问介词前面有无偏谓。介词的前面如果有偏谓（偏谓一般是非谓语动词），应为状语，修饰该偏谓。如果介词与偏谓之间还有名词，则需根据其他条件辨别它是名词的定语，还是偏谓的状语。介词短语位于偏谓后，处于谓语前，修饰偏谓而不修饰谓语。介词短语位于偏谓后，同时还处于谓语后，一般也只修饰谓语。例如：Next came the idea of depositing these circuits by evaporating the metal at high temperature in a vacuum. （紧接着出现了通过在真空中高温蒸发金属来沉积这些电路的想法。）

第七，查问介词与其前后有无语义联系。当 NPANP（介词在两个名词词组之间）的结构出现在谓语或偏谓后面的时候，往往给确定其句法功能和语义带来一定的困难。这时就必须通过查问语义联系的手段来确定。

关于语义联系，前面已经说过，这里再补充说上几句。"talk in English" 也可以说 speak、write、publish、report in English，但不能说 eat in English 或 drink in English，原因在于 talk、write, speak, report 等词与 English 有语义联系或语义搭配关系，而 eat drink 则无。

是不是可以说，in 见到 publish 就可以肯定地译成"用"呢？不行。in 后如果是 English、German、Chinese等语言名称，是可以这样译的；如是 England、China、Japan 等地点姓名词则应译"在"。从这里不难看出介词分析对语义联系的依赖性，以及对词（尤其是名词、动词）进行详细分类的必要性。

第八，查问介词与其前后其他词的语义搭配关系。进一步分析介词与其前后词的语义搭配关系，就会帮助我们确定介词短语的句法功能和意义。例如：

The fish was bought by the cook.

The fish was bought by the river.

这里，虽然两个 by phrase 都处于动词 bought 之后，都得到"前介状 A"，但前者当"由"讲，后者却作"在……旁边"讲，差别在于"cook"是有生命的名词，而"river"是地点性的名词。

总之，所谓图示处理，就是要从机器翻译的要求出发，把对介词进行分析研究的种种结果制成图，进而编成一定的程序，为机器加工介词提供有效的信息。这项工作做得越严密，越细致，机器理解和处理介词的人工智能程度也就会越高。

5.5 解释程序制导的表规则算法 在机器翻译中，语言规则即可以直接用程序语言的形式表述，又可以采取解释程序制导的表规则算法。在系统的研制试验阶段，规则的填加、修改极其频繁，如果规则与程序不致导致动规则即动程序的连锁反应的话。因此，在试验系统中采取解释程序控制的表规则算法可能是有益的。

为了便于表规则的建立、补充和修改，要求用较精密的算法研制有关的维护程序。规则表是一个相关文件，记录中的排序参数由名称、序号和增量组成。增补规则时借助于排序参数和排序中间文件即可实现表规则的序列更新。与规则表相匹配，还设有线加工文件，记载各线规则模块的起始相关地址。

表规则的记录格式参见表18。

右表列举的是有关名词组转换合成的规则片断。类型中的 F (Fixed) 表示定位取词，P (Perform) 表示执行一项操作，G (Go) 表示规则转向和循环检测。逻辑符提供"与"及"或"的关系，例如第 2 条规则中的逻辑"或" (R) 表示词类 (CL) 为冠词 (D) 或者副词 (F)。考虑到法语名词组的后置组分需要前移的特

表 18 表规则的记录格式

排序参数	层号	类型	A词	放过	B词	项目	算符	检测、执行内容	逻辑
N P 0 0 1 0 0	0 1	F	G C			C L	=	N	
N P 0 0 2 0 0	0 2	F	H 1			C L	=	D F	R
N P 0 0 3 0 0	0 3	F	H 2		N S	C L	=	F A	R
N P 0 0 4 0 0	9 9	P	H 1		H 2			S Q T L	
N P 0 0 5 0 0	9 9	G						0 0 2	
N P 0 0 6 0 0	0 2	F	H 1			C L	=	A	
N P 0 0 7 0 0	9 9	P	H 1		G C			F W P N	
N P 0 0 8 0 0	0 2	F	H 2			C L	=	A	
N P 0 0 9 0 0	9 9	P	H 1		H 2			S Q T L	
N P 0 1 0 0 0	9 9	P	H 2		G C			F W P N	
N P 0 1 1 0 0	0 2	F	Q 1			C L	=	D A	R
N P 0 1 2 0 0	9 9	P	Q 1		G C			S Q T L	
N P 0 1 3 0 0	9 9	G						0 1 1	
P P 0 1 4 0 0	0 1	F	G C			C L	=	V	

点，所执行的操作除顺序同号 (SQTL) 外，还包括前位同号 (EWPN)。由于采用了规则转向和逻辑处理等作法，规则数目可以得到精减。

主控程序将中间结果文件逐句送句子场进行规则检测和信息更新。在工作区尚须设置规则场和规则站。例如，在加工名词组时，可按线加工文件提供的起始地址把表规则文件的有关记录读入规则场。当加工词为名词时开始进行规则检测。与主程序相匹配并被它调用的有一系列子程序，诸如：扫描流程、放过查找、规则检测、规则操作、格式打印、译文综合，等等。子程序还可

以嵌套，例如译文输出子程序还可被译文综合子程序所调用。

表示系统总体流程的线扫描次序可设置为一个便于修改的相关文件，用以指示各线的加工序列、加工对象和扫描方向。这样做有助于实现总体流程的随机性和灵活性，从而便于对不同的线流程设置进行对比性的调试，并从中选择出较为优化的线加工序列。通过外汉题录机器翻译试验，我们认为，一种可行的步署可以是：第一线为名词组合成，第二线为动词、标点、连词分析，第三线为歧义、配价分析，第四线为介词分析。

考查一个现役词的静态或动态信息是否与一条规则的内涵限定相符，是借助于调用规则检测子程序来实现的。即使在定位取词的情况下，也必须考虑到可能有各种随机的和不确定的参数在表规则中出现。例如 A 词可能是该词 (GC)、后1词 (H_1)、后2词 (H_2)，也可能是前1词 (Q_1)、前2词 (Q_2)，等等。项目可能是词类 (CL)，也可能是成分 (CF)、词形 (CX) 或其他指标。此外，算符和逻辑符也可能呈现几种不同的变体。在这种情况下，就要求规则检测子程序对这些随机参数分门别类地进行归纳、概括和抽象，从而实现带有解释性质的规则检测。

被检测的每一条规则须从规则场调入规则站中。为了进行有效的解释，我们把规则站里 A 词的随机参数一律送入统一的变量“ZC”（找到词）中去。如果经过核实，认定被检测的项目是词类，算符为等式，逻辑符为“或”，那么，我们就可以借助类似下面的 COBOL 程序语句实现规则检测了：

```
IF L (2C) = ST-NR(1) OR = ST-NR(2) GO TO
JL-Y ELSE GO TO JL-N
```

这里的 ST-NR 是规则站里存放检测内容的区域，尚可进一步划分为若干小区，以便进行或及时的处理。通过对一条规则的检测，如果规则限定与现役词信息相符则为是取，不符则为否取。在此基础之上，主程序即可决定从规则场中再取出哪条规则置入规则站，以便对现役词施行进一步的检测。目前我们贯彻的

基本原则是：是取调入层号比该条更大的规则，否取则调入层号与该条相同的规则。一般地讲，在取规则的过程中，如果遇到层号比该条为小的时候，则为自然转出口。但是，在某种情况下，出于规则转向和循环转向的需要，如果在一条规则中特别标明“否取小号不转出”(NS)，则仍须将那条层号较小的规则调入规则站继续检测。

当通过一系列的检测达到99层规则时，则须调用规则操作子程序，根据执行内容实现对某项结论的操作。在执行完一条99层操作之后，如果不是规则转向或特别标明转出口信息的话，则应继续取下一条规则进行检测或操作。

99层的规则操作经常是在两个相关词之间进行的。与规则检测子程序的归纳法相类似，将规则站 A 词区中的变量都归纳为 ZC，将 B 词区中的变量都归纳为 BC。根据各种不同类型的执行内容，将“顺序同号”(SQ)、“顺序链接”(SL)、“前位同号”(FQ)、“前位链接”(FL) 等信息送 ZC 词的链特区。与此同时，一律以 BC 变量向 ZC 词的拉链区赋值。这种规则操作的程序实现手段可大致示意如下：

```
MOVE BC TO RT (ZC);
IF ST-NR = "SQTL" MOVE "SQ" TO LT (ZC).
```

由于名词组的合成及其等句素的词序调整和超句素的语序调整在法汉机器翻译系统中占有十分重要的地位，(诚然，在实际方案中，最有必要采取表规则的并不是 NP 合成规则，倒是介词分析最有必要采用)，下面就结合这一问题介绍表规则的制导和有关子程序的调用。

请看下列两条题录：

1	2	3	4	5	6
D	N	P	N	A_1	A_2

(1) L' Acceptabilité des Formes Verbales Swrcomposées

(动词超复合形式的适用性)

1	2	3	4	5	6	7	8
D_1	N	D_2	F	A	P	D	N

(2) La Particularité la plus Remarquable de la Planète

(大行星最显著的特征)

我们看到，这两条题录中出现了三种结构类型的名词组分布式：DN, NA_1A_2，D_1ND_2FA。我们还看到，根据译文生成的要求，第二种分布式须转换为 A_1A_2N，第三种转换为 D_1D_2FAN。

按照目前题录翻译线加工流程的规划，第一线即为名词组的加工。扫描方向由句首向句末逐个取词，若不是名词则须放过，若是名词则调用名词组合成规则进行检测。由于采纳了这种词类驱动法，因此各线规则模块的第一条规则必定检测为真。实际上，只有这种首条规则才具备该规则模块中唯一的01层层号。在上面引述的第一条题录中，第2号词 Acceptabilité 和第4号词 Formes 为名词类驱动词。

首先对第2号词进行规则检测。第1条规则是取，增加层号找到第2条规则，其后一词既不是冠词也不是副词，否取找到下一个层号也是02的第6条规则，后一词也不是形容词，否取找到第8条，再否取找到第11条，这时前一词为冠词，是取按第12条99层规则调子程序取结论，组成名词组 DN，继续运行，转出口。

再看第4号词的规则检测。调入第2条规则，否取找到第6条规则，是取按第7条规则进行操作，之后继续进行第8条规则的检测，是取连续进行第9、10两条的规则操作，得出 A_2 给 A_1 同号并移位至 N 前给同号的信息。

下面，我们再来考察第二条题录中第2号词 Particularité 的加工流程。

取到第2号词进入第2条规则进行检测，是取将第3条规则送站检测，仍为是取，进行第4条规则的操作；F 给 D_2 同号，按第5条规则的指令，转向第2条规则。继续检测之前需重新定

位，原来的 H_1 被紧缩化掉，原来的 H_2 变成了 H_1。由于此时的 H_1 为副词 F，第 2 条规则检测的结果为是取，找到第 3 条规则，由于 H_2 词是形容词 A，仍为是取，执行第 4 条规则的操作，令 A 给 F 同号，再次转向第 2 条规则，经过重新定位，此时的 H_1 已经是 A，故检测为否取，找到第 6 条规则，检测为是取，执行第 7 条规则的前位同号的操作。继续执行第 8 条规则，否取转第11条规则，Q_1 为 D，是取进行第12条规则的操作，N 给 D_1 同号，再按第 13 条规则的指向转第11条规则，因又经重新定位，故否取至第14条规则，层号较小，转出口。

在第一线加工中，对于每个句子中的非本线类驱动词也不妨作出某种记录。如果尚有下一线的类驱动词，那么有关的句子还应接受另一组规则模块的检测；如果根本没有，当然也就可以免去那一线的扫描了。一组题录通过各线加工，已经具备较充分的链特征和拉链信息。这时，即可仿照多叉树链结构递归拉链的做法进行调序了。这项任务借助于调用译文综合子程序来得以实现。

机器翻译的发展，同任何其他事物一样，也是由简单到复杂，由初级到高级循序渐进的。有人模仿计算机的分类，把机器翻译分作三代：词对词翻译是第一代，具有语法分析和句法分析能力的是第二代，具有语义分析能力的是第三代。这个发展过程，正好反映了对翻译问题的复杂性逐步加深认识的过程，同时也反映了机器翻译研究不断提高和机器翻译系统不断完善的过程。根据目前发展的情况来看，实用性机器翻译系统的出现，不是在遥远的未来，而是指日可待了。根据学者们的一般估计，机译产品将在几年之内流行于世，这是完全可能的。但是，哪个国家先实现这个目标，一个国家何时实现这个目标，这就要看各个国家的努力了。我国的机器翻译专业人员应当加倍努力，积极创造条件，促进机器翻译在我国早日实现，使它成为加速实现我国四个现代化的一个重要工具。

6. 情报检索

6.1 情报检索 知识信息或情报按照特定方式贮存和按照特定需要查找的过程即是情报检索 (information retrieval 简称 IR)。采用电子计算机实现这一过程即是计算机情报检索。

早期的情报检索是手工检索(包括穿孔卡片)，后来发展成机械检索(包括机电检索系统、光电检索系统)。20世纪50年代初期出现了计算机情报检索系统。近年来，情报检索出现崭新的面貌，进入了一个“情报—计算机—电讯(包括卫星通讯)”三位一体的新时期。

计算机情报检索(以下简称情报检索)，按存贮情报内容的表现形式可分为以下三种：

1 数据检索。存贮的信息是数据，检索时要查询数据资料档，并对提问输出答案。

2 事实检索。存贮的信息是各种事实，检索时可以对被检索的事实作某种逻辑推理，并进行比较和分析，然后再输出答案。

3 文献检索。存贮的信息是文章标题、著条项目和关键词组成的文献单元，检索时按提问检索词查询文献资料档，然后输出文献题录和文摘。

按存贮情报的时间，可分为：

1. 现刊检索。检索时可提供当前刊物上的情报。

2. 追溯检索。检索时可追溯若干年前的情报。

按计算机检索的方式，又可分为：

1．脱机检索。以计算机批处理方式进行检索。

2．联机检索。检索时利用计算机近程或远程终端进行人机操作。

美国是最早实现计算机情报检索的国家。1954年，美国海军军械试验站图书馆利用 IBM-701 电子计算机建立了世界上第一个计算机情报检索系统。世界上最早的联机情报检索系统也是美国研制的。目前，比较出名的联机检索系统有 ORBIT、MEDLINE、DIALOG 和 ESA-IRS 等系统。

我国从1963年开始进行机械情报检索的研究工作，1965年进行了试验。70年代以来，开始研究计算机情报检索，1975年首次进行了计算机情报检索试验，1977年又进行了联机检索试验。1933年在中国科学技术情报所建立了连接美国和欧洲主要国家的数据库联机检索系统。这个系统通过意大利的 ITALCABLE 分组交换中心与欧洲空间组织的 ESA-IRS 系统连接，并由数据交换网转接到美国的 DIALOG、ORBIT 系统。这样，我国就可以在北京利用通讯卫星检索到欧美200多个数据库的文献。

6.2 情报的存贮与检索过程 上面已经谈到，情报检索由存贮和检索两大部分组成。它们是既对立又统一，相辅相成的两个方面。情报检索的全过程可用流程图示意如下：

①提取二次情报的标引加工⟶②在计算机的外存贮器上建立数据库的文档存贮(造库)⟶③用户按自己的需求列出提问式进行再标引⟶④按提问标记在文档中查找有关的纪录（查库）⟶⑤根据检索结果进一步借阅或复制原文。

由此可见，①与②是情报存贮(系统建立)，③到⑤为情报检索(用户服务)。②属于数据文件的建立，需由专门程序或软件得以实现，④属于数据文件的查找，亦需专门软件承担。①与③都面临一个共同的问题，即是“标引”(indexing)，文件的造、查都是建立在标引基础之上的。科学地加以标引，可使得文献的预制和日后的检索达到和谐与统一，否则的话就会造成存贮与检索的

冲突，影响检索效果。因此存贮标引应当在最大程度上附合检索标引的客观需要。情报检索系统必须在建立、存贮时尽量考虑到为用户日后的方便服务，做到查准、查全。由于标引这一概念十分重要，下面将再作具体阐述。

6.2.1 情报检索的建立——情报存贮 建立情报检索系统的先决条件是从一次情报中提取二次情报。这个过程与图书资料的分类登录颇为相似。所谓一次情报，即是指图书、报刊、杂志的原文，它们既是情报中心收藏的对象，又是情报需求者寻求的对象。但是在收藏与寻求之间不是以一次情报的形式直接进行的，而是借助于二次情报。所谓二次情报，一般是指文摘、索引、书目卡片等等。情报需求者首先要查阅二次情报，方能借阅一次情报。事实上，计算机情报检索与人工检索有着共同的基础，即是为提取二次情报所作的标引加工。计算机情报检索系统的文档相当于由众多书目卡片组成的卡片箱，而文档的一个记录正相当于一张卡片。在文档记录上存贮什么信息或在一张卡片上填写什么内容，这些也正是标引加工所要考虑的问题。

标引加工是对一份文献书目及内容的抽象、概括过程。所选择的登录项决定了实现这一过程的方式。对于计算机检索系统来说，这些登录项的选择决定了文件的记录格式，也称作文件的数据结构。通常的文献均可选取以下这样一些登录项：标题，作者，来源，出处，文种，馆藏单位，索取号，文题词，摘要，等等。下面列出一个记录的登录样品。

标　题：计算机情报检索系统
作　者：王景寅
来　源：自然杂志
出　处：VL_4；1978.8；P223
文　种：中文
馆　藏：情报研究所
索取号：J＊26853

主题词：计算机；情报检索

摘　要：介绍计算机情报检索系统的原理、历史与发展趋势。
(缩微胶卷号 TB8047，价格5.70)

检引方式不是唯一的，例如在上述记录中，可把“时间”从“出处”中指出来列为独立的一项。但最主要的区别在于对文献内容采取何种登录方式。上面介绍的是主题词检引法，另外一种方法是关键词检引法。这两种方法有一些共性，但也有区别，下面分别介绍一下这两种检引法。

关键词 (key word) 是从文献原文中挑选出来的，例如从论文的篇名或书名中选出最能表示主题的词语。将关键词用于标引词 (index word) 必须原样照搬，不得改变。因此，关键词也称为文献语言或自然语言。由于它们的出现形式完全是随机的、因人而异的，故形式多样，同一概念也可采取五花八门的表达方式。例如计算机也有人称为电子计算机、电算机或电脑等。如果一篇文章题为《电脑与情报检索》，那么，用关键词法标引的话，我们只能选择“电脑”作标引词。用这种方法进行建档标引显然是较为容易的，甚至可利用计算机自动地从文献中选取题内关键词 (KWIC 即 key word in content)。

上面讲过，建档标引是与检索标引相配套的。采取关键词法进行建档标引的系统，在检索标引时也必须采取“电脑”这一关键词才能查到《电脑与情报检索》这篇文章，如果采用同义不同形的“计算机”作检索标引词，则无法查到这篇文章，为了提高检索的查全率，建档标引可采取各专业的标准术语，这就是主题词标引法。主题词也叫“叙词” (descripter)，它们是从词义的关键词中归纳、提炼出来的，亦即规范化的关键词。例如，用“计算机”作为主题词，标引建档，实际上也包括了题名中虽然没有“计算机”这个词，但却有其同义词(“电脑”、“电算机”)的那些文献。于是，用主题词进行检索标引则不必进行同义词替换即可做到查全。但是，为了便于实现主题词标引法，有必要事先编制出一套《主题词表》，

作为存档与检索的标引工具。通过查主题词表，即可明确同义术语的替代关系及规范化形式，找到标准化的术语。由此可见，主题词表的建立同术语的规范化有着密切的联系。有关主题词表的问题将在6.3进行详细讨论。

从计算机文档库的建造的角度来看，必须把按给定数据结构经过标引加工的二次情报建立在计算机的外存贮器上。由于这个过程不是一次实现的，必须按记录逐一进行。通常的做法是在终端的键盘上将一个记录的内容输送到内存缓冲区中，然后再存入作为外存贮器的磁盘或磁带上。一个情报检索系统的文件档实际上是众多二次情报记录的集合。笼统地讲，也可以说是一个数据库（data bank）。与其他任何数据库一样，情报检索系统的数据库应当有完备的维护软件提供数据的建造、修改、嵌入与删除的方法。这样，不但能确保按照标引方案建造检索文档，还能随时做到数据更新。

按照文件的组织形式可分为顺序文件、相关文件和索引文件三类。顺序文件没有地址，必须从文件的开始按照记录原有顺序检索到所需的记录。因此，存取时间取决于检索记录在文件中的位置，越是靠前的越节省时间。相关文件则是经过按某一登录项排序的文样，检索时可按相关地址号进行二分法查找。索引文件的每一条记录都有相应的地址，而这种地址是由能唯一确定该记录的某些登录项充当的，诸如主题词和其他具有检索意义的标记等。采取索引文件可以按地址随机查到所需记录，不受存贮位置的影响。

6.2.2 情报检索系统的应用——情报查找 情报查找是从用户角度来探讨情报检索系统的。目前一般都采取联机对话式检索服务。系统可提供多样化的检索途径，形式菜单，供用户选择，例如：

（1）按照作者姓名检索；

（2）按照主题词检索；

（3）按照文献出处检索；

(4)按照综合条件、逻辑表达式检索。

用户根据自己的方便，可选定任何一种检索途径，按下相应的功能键，进入相应的工作状态。如果选定(1)，屏幕就显示出"请指定作者姓名"的字样，当用户给出作者姓名之后，屏幕上则显示出一条该作者的二次情报记录，如果回车，则出现下一条记录，直至结束。最后回到菜单状态。

如果选择主题词检索途径，最重要的问题在于使提问词和文档中使用的主题标引词取得一致。因此查看叙词表，明确哪些术语可以充当标引词，则是完全必要的。例如在以"计算机"（不用"电脑"）、"自行车"(不用"单车")、"船舶"(不用"轮船")为标引词的叙词表中，用户如果忽视了标引词的选择，就很可能检索不到所需的文献。

在情报检索中，有两种性质的误差。如果漏掉了检索者所需的文献，未能查全，为第一误差。如果用户在查到所需文献之外还查到了一些本来不需要的文献，即未能查准，则为第二误差。比较而言，第一误差是严重的，不能容许的。而第二误差则较为次要，但也应予以控制。归根结底，用户需要的是既要查全（无遗漏)，又要查准(无多余)。为此，根据实际需求，综合各种条件，用逻辑表达式进行标引检索是行之有效的。

逻辑表达式的运算符号包括以下一些：

AND: 逻辑乘，可以用"∧"或"&"表示(诸参数同时具备才为真)；

OR: 逻辑加，可以用"∨"或"＋"表示(二者具一即为真)；

NOT: 逻辑否，可以用"¬"或"～"表示；

FOR: 逻辑优加，可以用⊕表示，先于逻辑乘执行。

除以上逻辑符号之外，还采用下述的数学符号：

等于＝，大于＞，小于＜，小于或等于≤，不等式≠。

在通常的情况下，逻辑乘先于逻辑加，而逻辑优加又先于逻辑乘执行。有时为了改变原有的运算顺序，可以加括号予以特别

的规定。

参加逻辑表达式的号数可以是标引记录的各个数据项。最常用的是主题词、作者及出处等等。例如用 C 表示 COBOL 语言，用 B 表示 BASIC 语言，用 L 表示非数值运算，用 T 表示报表，如果用户想要找的主题是与非数值运算有关，但与报表无关的 COBOL 或 BASIC 语言的文献的话，则可用如下逻辑式表述这一题材 (S)：

$S=(CVB)\ \Lambda L\Lambda_{7}T$

由于逻辑优加先于逻辑乘，故亦可表述为：

$S=C\oplus B\Lambda L\Lambda_{7}T$

当用户的要求包括更多的综合限定时，应当首先将思路捋清，再用逻辑表达式正确地表达出来。例如某一用户希望得到满足如下条件的文献：

（1）发表在《航天技术》期刊上且具有主题词“火箭”或“燃料”的文章；

（2）作者张天奇的全部文章；

（3）文献发表时间均须在1982—1987年之间。

根据这三项要求，应当怎样列出逻辑表达式呢？首先让我们分析条件 1。两个主题词之间的关系是 OR，即逻辑加，它们与期刊的关系是 AND，即逻辑乘。所以条件 1 的逻辑式可以表述如下（用 S 表示主题词，用 O 表示出处）：

(S: 火箭 V S: 燃料) ΛO: 航天技术

下面再来看这三个条件之间的关系。很明显，条件 1 同 2 的关系是 OR，即逻辑加，因为用户的条件 1 并不是以条件 2 作为前提的，也就是说可以是任何作者的文章。条件 1 同 2 的逻辑表示如下(用 A 表示作者，为了避免逻辑加的括号，也可用逻辑优加)：

S: 火箭⊕ S: 燃料 ΛO: 航天技术 VA: 张天奇亦可缩略写成：$S_1\oplus S_2$: ΛOVA。

条件 3 对于条件 1 及条件 2 来说都起到限定作用，因而是逻辑乘。该用户的综合三个条件的完整的逻辑表达式如下（用 Y 表示年份 1982—1987）：

$(S_1 \oplus S_2 \Lambda O \Lambda A) \Lambda Y$

这个检索标引过程在计算机终端上是采取会话形式进行的。在初始菜单上选择逻辑表达方式之后，终端屏幕上即显示出如下提问形式：

请按照你的需求列出逻辑表达式

参　数	逻辑符
______	______
______	______
______	______
______	______
时间：______	文种：______

所谓参数即是任一标引登录项，由用户自行定义。如果是主题词则先打个 S/，再填上相应的主题词；如果是作者，则先打个 A/，再填上相应的作者名，等等。逻辑符即是 AND，OR，FOR 等等。由于时间与文种单独列出提问，它们与其它用户自定义的参数显然是逻辑乘的关系。

如果某用户的文献需求仍然是上述三个条件的话，须按如下形式在屏幕上键入：

参数		逻辑符
S/	火箭	POR
S/	燃料	AND
O/	航天技术	OR
A/	张天奇	
时间：1982—1987		文种：中文

按照上述形式标引之后，按下回车键，即可出现第一篇所需文献，再回车，再出现下一篇，如此反复，直至穷尽。文献显示

的形式恰是文档标引的一个记录，其格式也和6.2.1中介绍的登录样品一致。这样，就可以根据摘要选出自己所要借阅的文献，根据索取号或缩微胶卷号到书库提取。

6.3 文献的分类与主题词表 我们注意到，无论是情报人员登录文献，或是用户查找文献，标引都是十分重要的。然而，为了实现正确的标引，必须对各种文献进行正确的分类，在此基础之上，建立主题词表，恰当地解决概念之间的同义、相关、上位、下位等关系。

科学是一个体系，它是分层次的。作为概念来说，也是按照一定的结构关系组配起来的，下面我们试用层次的划分原则来表达一下最高层次至最低层次的分类原则。最高层次用01表示。层号越大，层次越低。层号相同者为同一层次。

01 科学
　02 人文科学
　02 自然科学
　　03 数学
　　　⋮
　　03 生物学
　　　04 植物学
　　　04 动物学
　　　　05 无脊椎动物
　　　　05 脊椎动物
　　　　　06 鱼类
　　　　　　⋮
　　　　　06 哺乳类
　　　　　　07 后兽亚纲
　　　　　　07 真兽亚纲
　　　　　　　08 有爪区
　　　　　　　08 无足区

08 猛兽有蹄区
09 食肉目
⋮
09 偶蹄目
10 野猪亚目
10 核脚亚目
10 反刍亚目
11 猛马象下目
11 真反刍下目
12 鹿科
12 长颈鹿科
12 牛科
13 羚羊
⋮

如果为一篇有关羚羊的文献进行分类，就要从01科学→02自然科学→03生物学→04动物学→05脊椎动物→06哺乳类→07真兽亚纲→08猛兽有蹄区→09偶蹄目→10反刍亚目→11真反刍下目→12牛科→13羚羊，一直细分下来，决定这种结构层次。

在文献的分类中，第2讲上位概念与下位概念。以上述分类表为例，高层次的概念为上位概念；低层次的概念为下位概念。当然，上、下位只能是相对的。例如"05脊椎动物"，对于"06鱼类"而言是上位概念，对于"04动物学"而言则是下位概念。由于上位概念嵌套下位概念，下位概念包含在上位概念之中，所以，上位概念也叫广义词，下位概念也叫狭义词。

主题词表也叫叙词表 (thesaurus)①，是建立在科学概念分类的基础之上的。编制主题词表的首要工作就是搜集各个专业领域文献中的常用术语。从词类来看，主题词主要是名词、形容词、

①为了提供标引和检索蓝本，1979年中国科学技术情报研究所主持编辑出版了十卷本的《汉语主题词表》(社会科学两卷，自然科学七卷、附录一卷)。

及其复合词。其次，为了使主题词规范化，必须从同义术语中选出一个。例如从“引擎”、“内燃机”中选出“内燃机”；从“电子计算机”、“电脑”、“计算机”、“电算机”中选出“计算机”等等。象这种有同义关系的术语。一经规范化，就只能利用固定的叙词标引，但在主题词表中，必须明确这种替代关系，以便于用户标引检索。除同义替代关系之外，在主题词表中还应当明确术语间的上、下位关系。为此，应当标明某一主题词上属什么概念，下分什么概念。在某些情况下，一些概念虽然从属关系不明确，但在意义上有部分重合，这时，就应当注明它们之间所存在的相互参照的关系，例如，“工艺”和“技术”，“振动”和“波动”就有这种参照关系。

综上所述，主题词表应包括：用、代、属、分、参几种格式。所谓“用”即是规范的主题词检索，用以替代那些不规范的同义术语。“代”即是同义替代，在这里应当注明那些不规范的同义术语。“属”即是属于何种上位词。“分”即是分出何种下位词。“参”即是相互参照的主题词。下面以“计算机”为例说明主题词表的这种格式：

电脑

　用：计算机

计算机

　代：电脑、电算机、电子计算机

　属：电子机械

　分：存贮器、运算器……

　参：电子计算器

电算机

　用：计算机

电子计算机

　用：计算机

由于主题词表很好地处理了同义词及概念从属关系，因而有

利于用户查准、查全。例如用上位概念的标引词可以查到下位概念的文献，用标准叙词也能查到以其他同义术语命名的文献。

情报检索的研究在我国发展很快。随着计算机中文信息处理水平的提高，中文情报检索已获得了很大的成功。用户只要有一些基本的汉字输入训练即可上机进行对话式检索。目前，有的部门甚至研制了中西文兼容的系统，这显然是十分有意义的。因为在很多情况下，用户关心的是某一主题的文献，而不愿受到文种的限制。在这种情况下，“计算机”与“computer”应视为同一检索词。为了做到中西兼容，建立多语种的互译词典则是非常必要的。

6.4 结语 语言是信息最重要的载体，是科技情报的主要负荷者。因此，文献语言研究的深度对于情报检索的效率有很大的影响。从这个角度上说，情报检索系统中的关键问题是情报检索语言的建立。这种语言应具备能精确表达文献主题和提问主题所需要的词汇和语法手段，不应产生歧义，不受用户主观因素的影响，并且便于用程序运算方式进行检索。为了提高情报检索系统的效能，获得较高的查全率和查准率，在词汇方面如何消除术语的同义性和多义性，在语法方面如何求得既经济又足够的语法手段来表达必要的语法关系，这些问题都还需要进行认真的研究。目前，计算机情报检索一般采用逻辑式提问，这对一般用户来说颇为不便，因为他们不熟悉逻辑式的表达方式。因此，如何过渡到利用自然语言提问的问题已经提到日程上来了。语言学工作者在这方面可以而且应该发挥自己的作用。

7．言语统计

7.1 言语统计的意义 言语统计除了与语言教学(编课本)、文字改革(简化汉字的设计)和通讯工程等有极其密切的关系之外，对汉字信息处理也有重大关系。音位、音节、声调等语音特征的统计研究，对以字音为主的汉字编码方案的设计(如是否标调,如何标调),对于文字和代码中字母的配置，以及对于打字机键盘的设计都有重要意义。另一方面，字元和字的频率统计，对以字形为主的编码方案的设计，对于键盘的设计都有重要的参考价值。

词的频率统计对于词汇码的确定，也有很大参考价值。当然，字频统计对于确定内部标准码更是绝对必要的。我们的汉字编码字符集就是言语统计的直接产物。言语统计的内容、意义还很多，例如言语的概率性对模式识别就有很大作用。

进行言语统计，目的在于根据量的描述给出质的评价，即依靠定量分析得出定性分析。统计结果一般是做出各种频度表，供各个不同专业的人员使用。

人工统计的历史很长，从陈鹤桥的汉字统计，到目前为止，主要是进行字频统计，而且主要是人工统计。近年来，利用电子计算机进行言语统计工作，既快又准，统计量不受限制，而且能提供多种参数，因而促进了统计语言学的发展。

7.2 言语统计的范围 统计范围是由任务决定的。最初的统计，主要限于词汇方面。例如，世界上第一部频度词典《德语频度词典》(1898年) 是德国人 F.W.Kaeding 编制的，他为了创造新速

记法而进行了大量词汇统计。又如，我国第一部字频统计的著作是陈鹤琴为了教学目的而编的《语体文应用字汇》(1925年)。随着科学技术和文化教育的发展，言语统计范围不断扩大。例如，为了解决通讯工程中提高清晰度和压缩言语信号的问题。进行了大量语音参数的统计；为了对语法进行精确描写，统计了语法形式、单位和模型的用法；为了找到各种功能风格以及不同作家的风格特点，开展了数量风格学的研究；为了求出能区分不同语言的计量指数，开展了类型统计研究；为了求出能确定亲属语言发生年代的统计方法，开展了语言年代学研究。此外。有人利用言语统计的数据进行密码破译工作和辨认匿名作品的研究；还有人根据信息论原理来研究语言的熵和羡余度。

最近几年，为了确定标准汉字表和计算机内部标准码，进行了大量汉字字频统计；为了满足语言教学和建立数据库的需要，进行了大量汉语词频统计工作。

7.3 言语统计中应注意的几个问题

(1)抽样问题。进行统计，除特定任务（如对某本书进行全面统计分析）外，一般都有抽样或选材问题。抽样不当就会引起较大的误差。选材的比例有很大关系，如医学材料占1/3，则整个统计一定带有浓厚的医学色彩。法国学者 C. Muller 认为，“频率概念如果不马上和分布率相结合，那频率概念的价值是很低的”。这是经验之谈。因此，抽样时要考虑均匀分布问题。人们一般都从政论、报刊、文艺和科技四个方面来照顾这个问题。当然其中还可分小类，如文艺中又可分散文和戏剧等。另外，年代和地域上的不同，也有重要影响，如“文革”时期，“级”、“阶”、“命”、“革”四个字，在新闻通讯中的出现率名列前茅，分别占第18、23、36、37位。也就是说，位于最常用的100个字之列。详见下页表格。

另外，还有一个抽样多少才算合理的问题。一般认为，一次抽样以一百万字(词)为宜。抽样太少，不足以说明问题。

频率高度 单字 / 体裁	阶	级	革	命
政治理论	18	19	32	34
新闻报道	23	18	87	36
综　　合	43	39	61	59

（2）要保持统计语料的完整性。一般人对科技著作、文学作品或报刊文章进行统计时，常忽略统计标点符号。他们认为这些符号出现频率高且无实际意义。其实不然，标点符号既能给出句长的数据，又能对鉴定不同文体提供参考。另外，象“高高兴兴”“干干净净”这类词语应单独统计，而不应人为地与“高兴”“干净”合并。如果合并，重叠形式的使用度就不见了。

（3）一般频度词表所含的内容：

1)词的频度：该词在语料中的出现次数。

2)相对频度：该词在统计出来的全部词中所占的百分比。

3)累积频度：对所列各词出现次数的累加数。

4)累积的相对频度：对所列各词相对频度的累加数。

5)分布范围及分布频度：即按不同风格或其他标准分别统计的篇章数及出现次数。

6)词级：把频度相等的词归为一级，频度最高者为第一级，依次排列，最低者(出现一次的)为最末一级。

以上是比较常见的内容。有的频度词表还有累积词数的频度及其相对频度，有的词表带有根据 $U=F\times D/100$ 公式算出的词的使用度，还有的词表带有根据 $I=a\log F+b\log Dt+C\sqrt{Ds\text{-}1}$ 公式算出的选词指数(这种指数用作校正选词误差的手段)。

（4）注意静态、动态之分。进行统计分析时，应注意静态数据和动态数据的关系。二者之间的关系有时是一致的，有时呈相反的趋向。例如，汉语单音节词在词典中占很小一部分（两千多个），在具有5万词条的词典中与双音节词的比例大约是1:16，但在文章中，单音节词的出现率相当高，有时在比例上甚至还要超过双音节词，原因是很多单音节词是最常用的词。因此，我们不能简单地根据静态数据来论述词的切分问题，即单单依据双音节词在词典中占绝大多数这一点来论证首先切分双音节词的合理性；这里要重视动态数据，因为切分问题主要涉及的是文章中的词。

7.4 语料库的设计　语料库是语言统计的基础。要统计，就要有素材，这些素材的合理组织，便构成了语料库。语料库又是言语统计的副产品。每次统计之后，原材料及所得的统计数据都保存在计算机中，这样语料库就可不断扩大。语料库编排得好，就能提供不少副产品，比如，统计第二本书之后，就可拿后一本书的数据同前一本书的数据进行对比研究。又如可以利用所存的语料编制逐词索引（concordance）和试题库。语料库编排合理与否，直接影响到语料效能的发挥和副产品的多少。

语料的标记是最重要的问题，所谓种瓜得瓜，种豆得豆，在这里能充分得到体现。如想得到词类参数，就得给每个词注上词类标志。如果词类参数得不到，其他依靠词类才能得到的语法参数就更谈不上了。标记词类可以不受研究深度的限制，有些词的分类，尚无定论也没有什么关系，可以如实地把各种可能性都标记出来，如一个词既可能是V，又可能是N，还可能是A，那就把它归为VNA类词。通过这种统计，有可能促进词类问题的解决。

7.5 中文信息统计的特殊性

7.5.1 形体信息和语音信息分而治之　中文不同于拼音文字，形体信息和语音信息泾渭分明，统计形体的频度可以利用汉字，也

可以用拼音(当然要解决好同音问题)。但是，若要统计语音特性的频，必须采用拼音方式。例如，统计声韵调，又如区分多音词“长 (cháng)、长 (zhǎng)”，“行 (xíng)、行 (háng)”，“和(hé)、和 (hè)、和 (huó)、和 (huò)”，等等。因此说，除用大键盘整字输入外，都有一个采用何种编码输入资料的问题。而拼音文字则无这一问题，用打字键盘直接往计算机内输入便可。这是中文信息统计的第一个难点。

7.5.2 词的划分问题 中文信息统计的第二个难点是如何划分词。这个问题在我国语言学中还未完全解决。在连续的汉字序列中，如何把词的界限划分开来，目前还没有一个现成的、明确的标准可以依靠。这一点恐怕是多年来为什么我们只进行字频统计而未进行词频统计的重要原因。同时，这一点也给中文信息统计以及其他中文信息处理工作造成了严重的障碍。可以毫不夸张地说，离开了词，寸步难行，因为任何一个信息处理系统都是以词为基础的。划分词的困难，在于不容易得到一个简明的、合理的、首尾一贯的方法。如果划分标准不同，统计结果也会不一样，划分不合理，矛盾重重，统计结果不会科学。

台湾出版的《常用中文词的出现次数》(1975)给词下的定义是：词可以说是句子的最小意义单位。但我们在这本书的词例中发现，“奋勇”是一个词，“抵抗”是一个词，而“奋勇抵抗”也是一个词。再有，“纺织”是一个词，“公司”是一个词，“纺织公司”也是一个词。还有，“研究”是一个词，“研究出”是一个词，“研究出来”也是一个词。究竟符合不符合词的定义呢？要知道，我们也会碰到类似的问题。标准不一样，统计结果也会不一样，所以，怎么划分词，还是大有文章可做的。

利用电子计算机进行词频统计，除了如何切分词以外，还产生了一个由谁来切的问题，由人切，还是由机器自己切。到目前为止，虽然提出过不少自动切分方案，如“最大匹配法”“逆向匹配”“双向匠配”“有穷多层列举自动分词法”“逆向扫描”“二字前

进、最长组配、试探反馈的自动词法"，但都还不是理想的办法，不是有较大的局限性，就是规则繁复，而且也未能彻底解决问题。因此，目前进行的词频统计都还是以人工切词为基础的。

问题的严重性还在于，搞自动切分，既影响中文信息加工的速度，又浪费机器许多宝贵的时间和空间，因为无论是哪种切分法，都需要配备一个相当规模的词表，都需要相当繁复的查询。从这一点来说，我们的文字信息处理要比外国的文字信息处理多了一道徒劳无功的操作。顺便再提一句，从整个语文信息处理来说，我们的无谓牺牲太多了，除了词的切分，还有烦琐的汉字编码，多种代码的机内转换，庞大的汉字库，以及高出人家好多倍的汉字点阵。这些都给我们中文信息处理工作的发展带来了严重影响。如果汉语拼音化早日到来，将会给中文信息处理带来革命性的变化，这是确凿无疑的。

7.5.3 区分同音词　中文信息统计的第三个难点是如何区分同音词。汉语中同音现象十分严重。据一份统计材料说，区分声调的话，45200个词中有5249个同音词，占11.6%；不分声调的话，45200个词中有17435个同音词，占38.6%。看来这个数据不可忽视。声调是汉语固有的特征，有巨大的区分能力。标调之后，还有11.6%的同音词怎么办呢？在口语中没有多少关系，我们说的话都是标调的，有说话的上下文，一般没有问题，只是偶尔对个别的词(字)需要加以解释，如谈到章太炎，怕学生记错笔记，说这个"章"是"立早章"，而不是"弓长张"。

但是，在言语统计中必须把这11.6%解决好。任何的同音同码都不能允许存在，否则，统计出来的数据是错误的。如何解决？现有的编码方案中，一般是依靠偏旁，也有应用其他方法的。如何利用偏旁，也有多种方式。不过，都还没有找到理想的办法。

区分同音词的问题不仅是中文信息处理的问题，同时也是文字改革的大问题。从这里我们不难看出，中文信息处理与文字改

革有着极其密切的相辅相成的关系。中国文字改革委员会正在进行的“四定”(定形、定量、定音、定序)工作，尤其是定量工作，有赖于中文信息各方面的统计研究。

7.5.4 统计内容较为广泛　中文信息可以统计的内容比较多。拼音文字一般统计词、音节、字母、元音、辅音及其搭配（包括词组）的频率、词长、句长，也就差不多了。汉语除上述这些之外，统计的内容还要广泛些：在语音方面还要统计声母、韵母、声调(包括轻节)及其分布，以及儿化、双声、叠韵等；形体方面还要统计偏旁部首、笔划及其结构的出现频率。

统计内容和统计方法(包括编码方案和存储方法)有密切的关系。例如，倘若采用拼音方式，就可以多得一些数据（如元音字母、辅音字母、声、韵、调等），但如果采用双拼或双打，则字母的统计就无法得到了。再如，如果只解决同音词而不解决同音字的问题，那么汉字的统计也就无着落了。如果把输入的东西变为数据库、语料库，那还可以得到许多副产品，如各种索引。

通过间频统计得出汉字的频度表，对汉字信息处理有重大意义，比如说进一步订正我们的字符集。

词频统计，对于其他方面，比如教学，有更大的意义。汉字频度表只能对汉字教学提供方便；至于对汉字的组合能力，即构词能力，构成哪些词，主要是作词头还是作词尾，等等，则无能为力。汉语中单音节词占多少，双音节词占多少，三音节、四音节词占多少，尚无比较可靠的材料。词频统计做好了，对认清汉语的面目将有巨大作用。无疑地，对拼音文字的设计，如词长多少为合理，也有重要参考价值。

8. 计算机辅助语言教学

语言教学是语言学最早应用的领域。因此，狭义的应用语言学专指语言教学。

语言教学主要涉及到人对语言的习得与教授的手段与方法，与人的心理活动相关。语言是一种社会现象。与社会的诸方面都有关系；因此，语言的教学研究与普通语言学、心理语言学以及社会语言学关系极为紧密。在我们组织现代语言教学的今天，应当从诸学科的综合应用的角度出发来探讨语言教学的方法和教材的编选。

计算机辅助教学是指采用电子计算机协助教学活动。从大的方面可分为面向教师的和面向学生的两个方面。具体地说就是请计算机协助教师从事诸如拟定测试题目这样花费教师精力颇多的一类工作，以及协助学生进行以人机对话为主的自习和练习活动。虽然计算机辅助教学涉及的教学学科众多，但本文探讨的主要还是语言教学。由此看来，语言教学和计算机辅助语言教学的主要区别点，就在于是人还是机器充当教学的组织者。这两者都是应用语言学的研究对象，只是计算机辅助教学还涉及到计算语言学这样一门边缘学科。

本章准备从以下几方面介绍计算机辅助教学。首先谈一谈什么是 CAI，然后再谈教学软件的实例介绍，最后再总结地讲述一下 CAI 在教学改革中的意义。

8.1. CAI 首先，从字面上来看，它是计算机辅助教学的英文

缩写，这个术语的英文原名为 computer-assisted instruction。顾名思义，就是采用电子计算机从事教学活动。为达到这一目的，须将预定的教学计划及内容程序化，从而实现计算机的课堂教学或辅助性的课外操练。将计算机应用于教学活动的想法始于1955年，此后陆续在许多学科的教学活动中进行了尝试。1965年以后，人们正式承认了计算机在教学活动中的作用。在各个科目的计算机辅助教学当中，语言教学尤为引人注目，因此人们又单独提出了计算机辅助语言教学这一术语，习惯上缩略为 CALI (computer-aided language instruction) 或 CALL (computer-assisted language learning)。将计算机引入到语言教学当中产生了巨大的积极作用，它给语言教学这一应用语言学的前沿领域带来了自动化和现代化的生机。事实上，国内外 CAI 或 CALI 的经验都已经证明：计算机是一种非常得力的语言教学的培训工具。

8.2.教学软件的种类 语言教学软件从服务对象上来分，可分为面向教师和面向学生的两大类。面向教师的属于工具型，面向学生的还可分为讲授型、操练型和模拟型三个子类型。从硬设备上来分，CAI 尚可将大中小型机算作一类，PC 机算作另一类。

工具型的软件是为教师服务的。首先，它可以给教师提供用于软件编著的专用程序语言，借助于这种编著语言，可以研制出各种面向学生的教学软件。另外，工具型软件还可以充当教师的智能助手，帮助教师自动地统计词汇，分析句型，拟定试题，编选书目索引，统计考试成绩，等等。下面以英语试题生成为例介绍一下这种面向教师的工具型软件。

英语测试自动出题系统实际上包括有英语试题源库、英语试题初选程序、英语试题标准库以及英语试题生成程序几大部分。

英语试题源库实际上是一个语料库。用 dBASE Ⅱ 软件建立。从严格意义上说，试题源库里装的并不是试题，而是难度与试题相当的大量英文资料，具体地说，就是大量的英文句子。试题源库由长度为 180 字节的记录构成。

试题初选程序是在试题源库中按给定的信息选择试题，工作效率比人工快千百倍。因此，试题初选程序是建立标准试题库的有力的工具。试题的语言现象由索引标志给出。索引标志可以是单一的词，也可以分离结构和句型。为了便于命题者进行筛选，还可以键入所需例句的需要量。例如给定的索引标志为 if … were，那么屏幕上将按需要量显示出含有这个结构的例句，诸如：

If it were not felt socially inferior to speak in a working-class accent, public schools would not be so much in demand.

由于试题源库的数据量极大，而且都是学到过的英语，按某一索引标志提取的句子在数量上和质量上是能令人满意的。

在试题初选程序加工的基础上，即可逐渐建立起标准试题库。根据试题的类型和格式，将试题分为四类子库。以每行60个字符计，每题不超过5行的试题有结构、词汇和用法的多种选择题，每题不超过10行的有改错题；每题不超过13行的有综合填充题；每题超过13行者有阅读理解和听力理解题。为了便于试题生成，标准题库中的试题都标有鉴别符，它指明试题的语法结构的属性和试题的内容分类及难度分级。此外，还须标有上述的格式符，以便在试题生成时作不同的打印、显示的格式化处理。

试题生成程序将各分项试题库的调用、索引、处理、打印都融合于一个总程序中，从试题库中提取相应的试题，并分门别类地按规定化的格式打印出一份完整的试卷，提取某一类试题时可以采取随机的方式，也可以允许命题者对检索到的试题在屏幕上进行筛选。试卷的布局均采取人机对话的方式，例如机器询问如下的问题，均应由命题人作逐一回答：

How many sections are there in your test?

回答试卷的分项数目。

Input the current titlename.

键入当前子试题库名。

Input the headline of the current section.

键入当前项标题。

How many kinds of items do you want for this section?

键入当前项试题种类数。

Input the sign of the current kind.

键入当前种类的索引标志。

How many items of the current kind do you need?

键入当前种类的试题数。

回答上述一系列试题后，程序开始进行。于是在屏幕上出现检索到的试题。如果命题人希望介入试题筛选工作，可以对检索到的试题进行取合。程序继续进行，直至生成一张完整的试卷。

以上介绍了面向教师的工具型软件，下面再介绍一下面向学生的软件。面向学生的软件有讲授型、操练型和模拟型三个子类型，但它们可以统称为“课件”(courseware)。

讲授型是通过屏幕课文显示，向学生讲授和解释课文要点，在这种情况下，计算机代替人向学生传授知识，即充当所谓机器老师。操练型是通过各种程序化的练习题，进行人机交互式的问答。学生作出回答后，机器可提供评价、修改和建议。学生的语言知识在操练过程中得到巩固和提高。模拟型是采用音响、画面和录象等技术模拟日常生活中的语言交际场合，使学生如同身历其境，诱导他们作出适当的语言反应。以上三种类型的课件实际上是互相贯穿的：操练型可能包含有模拟型，讲授型又可能包含有操练型。它们的目的是统一的，即帮助学生掌握语言、习得语言。下面介绍一下教件使用的某些具体情况。

外语教学课件从阅读、写作、听力机会话等角度为学生提供了习得与训练的机会。事实上，各种课程的上机都可借助于屏幕显示的操作信息而变得简单易行。例如在外语的听力和翻译的课件中，学生按下各种功能键即可获得有关词汇、词法、句法方面

的提示，或者得到有关错误的分析，查看正确的答案以及重新调入一段听力练习，等等。提示用的语言一般都采用学生的母语。下面分别介绍课件在阅读、写作和听力训练方面的应用。

语言阅读的基础是词汇教学。例如一个有关讲授房屋和家具名称的外语课件，学生在屏幕上首先看到一个自己的母语和一间没有家具的房间。课件要求学生把这个母语词译成某种外语，并打印出来。假如这个母语词是“床”，学生翻译准确无误，那么在这间房屋里就会有一张床显示出来。这种图像动画手段在词汇教学的 CAI 中已在广泛应用。在外语词汇教学的课件中，有一种母语和外语的词汇匹配练习，机器可以记录学生匹配所用的时间和达成正确匹配的比例，并可把参加这一练习的学生成绩作一个相互比较。

计算机在外语阅读理解练习中可提供十分理想的课件。这是一种操练型的课件。对时间的制约有两种方式，一是控时阅读，屏幕上以预先选定的速度显示阅读材料的文字，速度的控制由机器自动执行。二是计时阅读，屏幕上以书页的形式显示阅读材料的文字，阅读速度由学生随机掌握，读完一页按回车键翻页，整篇文章读完后，机器可显示出阅读所花费的总时间。阅读方式的选定可采取人机对话的形式。阅读过程中，有的系统可提供一部机器词典供学生查找生词。阅读以后，大部分课件都配有检查阅读理解的多项选择练习题，以学生键入 A、B、C、D 的方式进行回答。答题之后，机器自动评分，必要时还可给出正确答案。

以上介绍了课件的一些类型。除按服务对象进行教学软件的种类划分以外，还可以从计算机的硬件设备方面进行分类。

从计算机硬件设置方面谈，可分为终端类型和微处理机类型两种。在微处理机之前，计算机辅助教学软件都是在大型或中小型机的主机上进行的，这就是终端类型。学生上机学习或操练时，坐在主机串联出来的众多的终端跟前。每人一台终端，包括键盘和显示屏幕。终端对主机的依赖性很大，靠主机控制指挥。所有

的语言教学课程都存贮在主机内，主机可以同终端设置在同一校园内；如果是大型机的话，还可能离它的终端千里之遥。目前国外大型机的 CALI 系统可串联终端600台，小型机也可串联100多台。

除终端类型的教学系统以外，近年来还发展了微型 PC 机类型的教学系统。从构成部件上来看，微机系统包括主机、打字键盘、显示屏幕以及磁盘驱动器。语言教学材料通常是存贮在软盘内，磁盘驱动器将教学软件从软盘载入微机内存。微机教学系统是各自独立的。虽然它们也可以组成更大的网络系统，但每一台微机即是一个独立教学单位。

以上两种不同的硬件设置虽然都可以完成计算机辅助语言教学的任务，但其间的差别还是不小的。终端型的系统较为昂贵，安装和维修均需要较多的技术力量。微机型的系统则价格较廉，不需要什么技术配备即可由主管语言与教学的单位组建微机语言教学中心。然而，大、小型计算机的终端系统比微机型的系统工作效率高。这主要体现在终端系统可以在显示屏幕上演示更多的教学材料，也能更方便地记载学生上机学习的测试成绩。此外，终端系统所能提供的教学辅助手段更多一些，诸如在作练习的可向学生提供各类机器词典的咨询。然而，微机系统也有一些独特的优点，例如音响生成和彩色图像等功能是终端类型所不具备的。另外，微机的功能已日益增强，从而给微机辅助语言教学带来了新的生机。

8.3. CAI 的优点　计算机辅助语言教学有诸多的优点。首先是可以切实做到“因材施教”。搞过多年语言教学的人都有一个体会：学生程度不齐是最令人头痛的事。由于要求不一，很难适当地组织课堂教学。CALI 可以解决这个矛盾。学生自己上机，自己根据个人不同的情况决定进度，比如是复习前一课呢还是学习下一课，全由学生自己决定。事实上，根据课件的软件设计，每作完一项课堂教学或练习，学生都面临一些交叉选择。例如读完

一般阅读材料之后，学生可从如下的菜单中选择进度：

(1)你是否打算再读一遍这篇阅读材料？

(2)你是否打算作一些理解问答题？

(3)你想知道理解问答题的正确答案吗？

(4)你打算阅读另外一篇文章吗？

(5)我们是不是今天就到这儿？

通过如上这些交叉选择，学生可以有充分的自由定出步调，从而真正实现了因材施教。

还有，机器教师不会给学生造成心理上的压力，无论学生提出怎样的要求，机器也不会有不耐烦的感情流露，这样就十分有利于学生轻松愉快地进行学习和操练。有时，为了进一步增加模拟课堂教学的亲切感，还可以通过软件设计使计算机人格化。例如机器老师给定一个名字，同时还把学生的名字也加入到计算机的应答中去。于是学生可以在屏幕上读到这样的字句："Hello, Jack, I'm Mr. Johnson. Welcome to our fast reading training course. Are you ready? Let's begin now." 当学生答对了问题之后，计算机则加以鼓励，显示出："Well done, Jack, come on, please." 事实证明，亲切的感情交流对于语言习得是十分有益的。

由于课件的制作是建立在教师教学经验的基础之上的，因此可以去伪存真，去粗取精，博采众长，汇总各种成功和先进的教案教法，集各类教学精华于一堂。正是因为这样，教学活动可以从教室的现场活动中解放出来，成为学生可随时调用的、超时空的活动。从教师的角度来说，由于获得了工具型软件的支持，教师的备课、教学、测试、研究等活动的效率大为提高，同时教学资料和学生的档案也都能更加便利地得以积累和保存。

以上列举了计算机辅助语言教学的一些优点。人们对这些优点的肯定是经历过一段过程的。

目前，愿意尝试计算机辅助语言教学的人越来越多了，但是不了解或不熟悉计算机辅助教学的人仍不在少数。此外，对于是

否应用机器进行语言教学也有一些争论。有人从哲学观点出发，认为语言是人文学科，用机器教语言是没有希望的，还有人对程序化的教材——课件提出了一些批评。但是，总的来说，人们逐渐认识到计算机在语言教学的某些方面确实可以帮不少忙，起到很好的辅助作用。例如阅读和听力训练，由于个体之间的接受、理解能力相差很大，在传统的课堂语言教学中，教师只好采取一种阅读理解或听力训练的“适中”的进度，然而，这样就难于做到因材施教，程度低一些的学员总是反映进度太快，而程度高一些的学员又总是反映吃不饱，因此采取计算机控制的语言训练是解决因材施教的有效途径。另外，在完成语言课程的家庭作业以及自习方面，计算机的应用也可导致更加理想的效果。程序化语言教学的好处还在于，采取人机对话式的学习，学生可以立即得到有关练习答案的反馈信息；采取面向特定学生的教学，使课程的设置和进度自动调节到与该学生语言程度最为合适的档次；采取以学生为中心的教学，课程进度可以由学生自己调节，也可以根据他对问题回答的水平进行随机调节。考虑到计算机辅助教学的这些优点，人们现在对机器用于教学的非议已经越来越少了，一些批评和讨论的交点主要集中在教材的性能上。有些人自觉不自觉地将 CALI 同结构主义语言学的训练方法联系起来，这实际上可能是一种误会。目前，交际语言教学理论比较时兴，这已逐渐成为课件编制的语言学原则。

与 CALI 的优点相连系的一个问题是它的研制方面的特点。实际上，产生诸多优点与功能的课件乃是语言学家、语言教师、心理学家和计算机科学家紧密合作的产物。语言学家根据学科内容提出某一课程的教材。由语言教师指出学习重点和教学方法。心理学家则制定教学方案和评定学习效果的原则。然后由计算机科学家把上述材料编制成程序。经过反复调试、修改，即可成为投放技术市场的课件。

由于课件的研制是一个跨学科的智力合作问题，这就涉及到

一个合作方式问题。顾名思义，CALI 为计算机辅助语言教学，主要的智力因素可分为从事计算机软件的与从事语言教学的两大类。在某些情况下，这两种智力可以兼而有之，由一人承担。但在不少情况下仍须几方面人员的合作。首先是有经验的语言教师和教材编写者将其教材、教案、教法和练习方法总结提炼出来，并加以优化，然后由有经验的程序编制人员编写成能够在机器上运行的课件。当然，还有另一种途径，即有经验的语言教师或教材编写者同时又是有经验的程序编制人员。这样全面的人材较为难得，但其工作效率较高，可大大缩短课件的研制过程。就任何一个国家来说，课件的编制往往是分散在各地区，各院校进行的。这就产生了一个加强横向联系的问题，否则低水平上的重复劳动会导致时间和人力的浪费。在美国有一个叫作 CONDUIT 的结构，就是负责协调各种机器版本教件的研制工作的。我国也应该加强课件研制方面的协调工作。衡量一个课件是否成功，应当有一个客观的标准，这就是看它在语言教学实践中所起的作用。有这样的情况，有的课件能够在教学中起到良好的作用，但是从程序设计的角度来看可能不太优化，花样技巧可能不太多。从公允的角度来说，这样的课件仍不失为成功的产品。上面讲过，教件的编制是一项跨学科的工作，语言教师编写程序可能比专业软件人员逊色一些，因此不宜从程序方面求全责备。事实上，如果把评价课件的重心移到程序设计本身而忽略它在语言教学中所起的作用，那是不太公道的。与这个问题相关的是，有的机构在编制课件时把精力过份地集中在采用最先进设备和技术方面，亦即追求大、洋、全，结果使成本成倍地增加。这样实际上是阻碍了课件的普及和推广。正确的态度是从本单位的经济实力出发，按照教学的实际需要，研制切实可行的课件系统。

8.4. 美国 PLATO 系统简介 下面介绍一下美国厄巴纳-尚佩恩的伊利诺伊大学的 PLATO Ⅳ型系统的教件在外语教学中的应用。

PLATO Ⅳ型系统是由伊利诺伊大学计算机电化教育研究试验室为外语教学研制的。该系统于1973年付诸教学实践，是计算机辅助教学的先驱。PLATO 采取的硬件支持是两台大型主机：CYBER 6500和7300 及其存贮设备和学生用的分时终端。教学材料存贮在主机里，学生坐在自己的终端跟前，即可将所需的课程通过微波、电视波道或电话线调入终端。在他上机过程中，始终与主机接通，关机时，他的记录存入主机。主机中存有每个学生的学习档案，对每一课程或教学阶段的学习成绩都可以随时提供咨询。终端也可供教师使用，外语教师可以查看学生使用终端上课的次数，以及他在各个语法学习阶段的作业分数，错误特点，以便全面指导学生的外语学习。PLATO 的另一特点是：除人机对话之外，在使用终端的人们之间，不论其相距多远，也可以进行相互对话，其中包括学生和老师之间的对话。比如学生在上机作语法练习时遇到了困难，他可以停下来，把他的问题提交给老师，学生间的对话可用于会话练习和集体讨论。

PLATO 系统借助于自己特有的程序语言 TUTOR，以及512×512点阵的屏幕，相应的键盘可供产生多种字符集的外文，其中包括中文、日文、俄文、印地文、希伯来文等等。TUTOR语言和高精度显示屏幕还提供彩色图像。这对于外语教学起到了十分重要的作用。例如在英语作为外语的课件（ESL）及希伯来文的教件中，基本词汇的教学及测试都是通过图像进行的。在法语课件中也通常用图像提供会话背景，让学生进行会话或作文练习。

PLATO 系统用于外语教学的另一个重要特征，是除键盘输入之外，还允许学生在屏幕直接触接输入，这主要是借助于触板技术而得以实现的。屏幕触接输入用于键盘之外的字符输入，还用于多重选择式的练习，学生直接在屏幕上作出选择可大大加快人机对话的速度。

PLATO Ⅳ型系统有一种用于外语教学的音响设备。在这种

音响设备支持下的磁盘像录音磁带那样录制了很多音响，不同之处在于，磁盘上的各段音响可以随机调入，只需在键盘上或在屏幕上给出相应的信息即可。

伊利诺伊大学讲授的各种语言课程都有 PLATO 课件。但大多数语种的课件都是选修的。这就是说，课件的学习成绩不纳入学生在该语言学习中的总分数。例如法语和英语作为第二语言的教学中， PLATO 的使用仅作为一种课外练习活动，并不要求人人参加。但有意义的是，尽管这种选修课不纳入学分，大多数学生也都踊跃参加了 PLATO 的上机练习。由此可见，计算机在语言教学中确实起到了很好的辅助作用。

PLATO 系统有一种印地语的课件，采用编码系统即可将键盘上的英语字母转换成印地字母。在教学的初级阶段有专门的课件教授印地字母的书写编码系统。又如，在德语教件中，为了练习形态变化，在屏幕上显示出一个德语词和许多词尾，然后叫学生指出哪个词尾是正确的。在选择正确的语法形式的训练中，最常见的练习是填空。例如在英语作为外语的教学课件中，在讲动词 be 时，学生首先见到在屏幕上显示出 be 的各时态及人称的用法总结。然后出现若干句子，由学生填空。当学生键入动词 be 的一个形式时，它便出现在那个句子的空缺处。如果填充正确，光标转入下一个句子。在德语关系从句的填空训练中，屏幕上半部分有两个简单句，底部有许多关联词和一个逗号。学生只要在某个相关词上触接，那个词便跳到上面构成复合句，如果选择有误，学生将获得课件的反馈信息。在法语和西班牙语的课件中有许多翻译训练。例如，学生根据显示的英文句子译成法文，并把法文键入机器内，由机器进行译文质量判断。如果学生在终端显示屏幕上见到："I returned that money to you six months ago"，那么，他就应当把译成的法文 "Jo vous ai rendu cet argent il y a six mois" 键入机器中。还有一种转写练习在英语、西班牙语和法语课件中十分普遍。练习采取的格式如 Bill /

come / class / yesterday，要求学生转写为正确的句子：Bill came to class yesterday。还有一种写作练习要求学生把一系列打乱了自然序列的句子按照写作逻辑次序重新组织、排号，组成一篇短文。尽管可以说 PLATO 系统在写作训练上已经具有十分丰富的软件，但由于机器教学的限制，目前的软件还不能修改学生自由写作时出现的语法、词汇或修辞错误。

至于听说训练，前面讲过，由于机器带有特定的语音装置，学生可以方便地进行这方面的训练。例如在学习印地语的语音时，学生听到一个音节的发音，但在屏幕上出现两种拼音形式，要求学生判断哪种拼音和听到的音节相互吻合，并用触摸法作出选择。中文课件的听力训练也采取类似做法。例如学生在屏幕上见到许多用中文书写的数词，然后他听到一个中文读出的数词。他的任务是挑选出与听到的数词相同的书写形式。还有一种练习是要求学生根据听到的中文数词键入相应的数字。在许多外语教学的软件中有一种最常用的听力训练格式，学生先听一个句子，听多少遍可以用他自己控制，必要时，他还可以叫机器把听到的句子书写在屏幕上。此外，他还可以把自己的读音录制在机器中，与机器的标准读音进行对照。在大段会话的听力训练中，比如在俄语听力理解练习中，学生一边听到俄语会话，一边在屏幕上看到与对话情景吻合的录像。听完会话之后，在屏幕上打印出许多与会话有关的用俄语提出的问题，供学生阅读和回答。由此可见，课件的听力训练是卓有成效的，它可以起到单纯课堂教学难以产生的效果，而且又比语音实验室还略胜一筹。

以上通过读、写、听等方面的训练介绍了 PLATO Ⅳ型系统在语言教学中所采用的工作方式。现在，这些工作方式已有不少移植到微机系统上。需要指出的是，上面介绍的只是 PLATO 外语教件中最通用的基础技术手段，更加成熟和灵活的工作方式的设计正在层出不尽地涌现出来。

8.5 CAI 在我国 以上介绍了美国 PLATO 外语教学系统。

在我国，近年来 CALI 已有较大的发展。据不完全统计，研制单位包括清华大学、北京大学、上海交通大学、哈尔滨工业大学、华东师范大学、广州外语学院、华中工学院等。

他们的研究成果基本涉及到面向教师的工具型软件以及面向学生的三类课件。本章介绍工具型软件时谈到的自动出题测试系统，实际上援引了清华大学的例子。目前，上述许多院校都发展了诸如此类的出题系统，这类系统提高了教师备课、教学、研究工作的效率，把他们从繁复的事务性工作中解放出来，从而能够把精力集中在富有研究意义的课题中去。从课件的语种上来看，国内的 CALI 系统主要是用于英语教学。例如上面谈到的快速阅读系统即是哈尔滨工业大学为英语教学研制的一种课件。此外，哈尔滨工业大学还推出了一种微机控制的英语听力学习系统，这是一种听力操练型课件。具体做法是：把听力练习录制成磁带装在盒式录音机中，并且把有关听力的问题与答案作为数据文件存贮在计算机中，由工作程序协调视、听关系。以英语听力为例，当工作程序进行时，首先播放听力练习题目，同时把该题可供选择的答案 A、B、C、D 显示在屏幕上，学生须从中根据听力理解选出一个正确的答案，并把选定的那个字母通过键盘输入机器。如果选择是正确的，屏幕上就显示出肯定信息并转入下一道听力题，如果选错，屏幕上即给出否定信息。如果学生认为有必要再听一遍，亦可按键使录音机快速反转，重新播放。当然，这一系统的关键在于对录音机的性能选定以及与微机沟通信息的接口设计。

上海交通大学所研制的教件很有特色。一种是可供学生直接使用的操作系统，另一种是师生交互的操练系统。在后一种情况下，教件分为两类软盘，一种是教师磁盘，另一种是学生磁盘。外语教师先利用教师磁盘根据教案教法和教学进度把练习规格、题目及答案存入学生磁盘，再由学生上机操练。这种双盘制的好处在于便于教师指导和练习内容的更换。

最后，值得一提的是华东师范大学研制的 ETRA 系统，这是一种充当英语教学与研究助手的智能软件，这个系统对言语统计也有较为成熟的经验。

8.6 CAI 的意义 计算机辅助语言教学意义深远。我们认为，将计算机引入语言教学，可以使教学质量飞跃发展。多年以来，我国的教育界一直探索改革问题，以前的有些做法显然是失败了。现在是改革之际，当我们重新考虑教育改革的时候，要充分认识现代化的重要性。

计算机辅助语言教学有助于克服传统课堂教学的种种限制。首先，语言习得须以操练为主，即使通堂都作练习，人均时间仍十分有限；其次，课堂教学无法照顾到学生之间的个体差异，因而很难掌握教学进度，以致外语教学的计划教案往往反复无常、众口难调。另外，跟班学习，经常造成学生的心理压力，也可能影响学习效果。CALI 从根本上冲破了上述限制。学生使用课件进行学习，时间全由自己掌握，最大限度地提高了单位时间的信息获取量和操练次数。此外，课件可以因材施教地有针对性地帮助某一具体的学生指出进一步的学习方向，诸如某些练习答对了则转入下一项练习，否则复习某些知识并重复该项练习等等。此外，课件运行速度也可由学生自己选择，排除了为适应同班其他人所产生的进度上的不适应性。由于机器老师总是不厌其烦，学生的心理压力自然消除。

从根本上说，CALI 使外语教学从封闭型的课堂教育转变为开放型的社会教育，从集体听讲转变为个别讲授，从学校教育转变为个人自学。这三方面的战略性的转变能够使教育事业有效地应付它所面临的社会进步对教育质量和数量的巨大挑战，从而有助于实现当今语言教学方面的改革。

9. 语音信息处理

本章在语音信息处理这个总标题下面首先介绍一下有关语音学和实验语音学的基本概念，然后再谈谈有关言语合成和言语识别的若干理论和实践问题。

9.1 语音、语音学和实验语音学 世界上存在的声音很多，有风声、雷声、机器的轰鸣声、动物的叫声，既有乐器优美的旋律，也有各式各样的噪音。但这些声音显然都不是语音。所谓语音则是指人类语言的声音。人类的语音是凭借声音和意义的结合而达到直接交际的目的的。对于对话双方来说，也都是靠语音而达到信息传递和接收的。因此，我们将人类各种语言中的声音称为语音。

语音学则是研究语音的一门科学。它主要是研究语音的成分、结构和变化规律，并探讨相应的分析手段和研究方法。具体地说，语音学可以帮助人们掌握发音、听音、辨音和记音的技巧，探讨语音的类别、变化组合的方式和各种特定语言的语音体系。

语音学根据其研究对象和研究方法的不同，尚可分为若干学科。不分语种，研究一般语音构成规律的学科叫作“普通语音学”，结合某一种语言或方言的叫作“各别语音学”。按语种间有否亲族关系来分，比较诸亲属语言的语音发展演变规律的是“历史语音学”；比较无亲属关系的诸语言之间语音异同的则是“对比语音学”。不借助任何仪器设备研究语音的叫传统语音学，而所谓实验语音学则是采取特定的语音学仪器来考查、记录、分析及研究语音的物理现象和发音方法。实验语音学最初叫作“仪器语音学”，

采用的仪器包括研究语音产生的生理机制的医学仪器和分析、测量语音特征的物理仪器。近年来，采用电子计算机进行语音的信息处理是当代实验语音学的一个新潮流。由于有现代化的机器设备作为研究手段，实验语音学对于各类传统语音学的研究无疑有重要的促进作用。

实验语音学的任务主要是研究语音的产生、传播和接收这三个领域的问题。事实上，这三个方面正好反映了语音作为会话交际的过程，因而构成了一般所说的“言语链环”。首先，语音的产生取决于诸生理机制，例如神经系统、口腔的肌肉活动以及声带、声腔的发音动作等等。这种研究语音产生机制的学科叫作“生理语音学”，或“发音语音学”。其次，研究语音产生后在空气中传播的物理特性的叫“声学语音学”。所谓语音的物理特性包括人们所常说的语音四要素，即音色、音高、音强、音长，关于这四个要素下面还要讨论。最后，探讨语音是怎样被人们接收的听觉和理解过程的学科则是“心理语音学”，也可称为“感知语音学”或“听觉语音学”。由此可见，实验语音学研究的言语链环实际上包括有关语音的生理、物理及心理的语音特性分析。这些就构成了语音的物质基础。

9.2 实验语音学的语音的特征分析

9.2.1 语音的生理特性 语言是人类特有的社会行为。人与其他动物的重要区别之一，就在于人能够使用语言、产生语音进行交际。毫无疑问，这首先就在于人脑的发达程度优于其他动物，因为语音的产生是在大脑的支配下通过神经系统指挥发音器官得以实现的。

不同的语音取决于发音器官活动部位和方法的不同。人类的发音器官包括呼吸器官，喉头与声带，口腔、鼻腔和咽头三大部分。

呼吸器官包括肺和气管。肺有左右两个，是呼吸器官的中心，提供发音动力，吸气时肺部膨胀，胸腔扩大，呼气时肺部收缩，

胸腔缩小。人类主要是借助于呼气而发出语音的，吸气音只占极少数。

喉头位于气管和咽头之间，系发音的声源所在。喉头由甲状软骨、环状软骨和两块勺状软骨所组成。声带位于这几块软骨所构成的圆筒形之中。声带由两片有弹性的带状纤维质薄膜组成，能开能合，其通道叫作声门。肌肉牵动勺状软骨的开合，使声带松紧不一，导致声门的开闭状态各异。若有气流冲开闭合的声门即可发出元音或浊辅音。

口腔在喉头的上方，分上颚和下颚两大部分。上颚是固定的，下颚却可以自由活动。上颚包括上唇、上齿、上齿龈、硬颚、软颚和小舌。下颚包括下唇、下齿和舌头。

咽头位于喉头之上，并与之相通。此外，它在软腭控制下还可以上通鼻腔，中通口腔。

鼻腔在口腔的上面，它也是发音的共鸣器。鼻腔本身不能活动。当软颚堵住口腔的气流通过时，即可因鼻腔共鸣而发鼻音，堵住鼻腔的通道即可发出口音。如果软颚小舌悬在中间就可发出口鼻音，例如法语的［ɑ̃］、［ɔ̃］、［ɛ̃］等。

9.2.2 语音的物理特性　物质经振动发声，振荡了它周围的空气，形成声波，在空气中传播，使空气产生疏密相间的有规律的变化。于是产生了语音的物理特性。上面已经谈到，语音有四个要素，其中音色属于音质范畴，音高、音强和音长可统称为超音质特征。

（1）音色。一种乐器的声音之所以和另外一种乐器的声音不同，原因就在于它们的音色不同。同理，语音中［n］与［l］，［ai］与［ei］不同也在于音色各异。语音中的这些音色变化主要是由发音器官的形状与发音方法不同所致。正是借助于这些音色的区别，也得以形成语言中的诸多的元音及辅音音素。不同的音色波形是不同的。音色可以独立鉴定。即使我们采用不同的音高、音强和音长来读某一个元音时，也仍能同另一个元音相区

别。在各种音色中可划分乐音和噪音两大类。语音中凡振动声带的发音都是乐音，不振动声带的是噪音。

（2）音高。顾名思义，音高即声音的高低，它取决于声波振动的快慢。通常用表示每秒振动次数的频率单位赫兹来衡量。因此，音高也通常叫作音频。音高取决于声带的状况。声带长、声音低；声带短，声音高；声带松，声音低；声带紧，声音高。音高虽然属于超音质特征，但在汉语中，由于声调正是由音高体现，因而它在汉语语音系统中的地位相当突出，作用相当重大。从某种意义上说，其重要性并不在音色之下。

（3）音强。音强即声音的强弱、大小，一般也叫音量，它取决于声波振幅起伏度的大小。振幅大，声音强；振幅小，声音弱。音强和音高是两个不同的概念。男人讲话一般声音低而强，即音频低而振幅大；女人讲话则声音高而弱，即音频高而振幅小。汉语的声调属于音高的变化，而英语的重音则属于音强的变化。

（4）音长。音长即发音时间的长度，它取决于发音体振动的久暂。音长同声音的音质关系不大，但在有些语言中在音长上可形成对立，如英语的 [i] 和[i:]。

9.2.3 语音的心理特性　语音的发出是为了使会话对象接收感知而获得信息。这样就涉及到人耳的听觉和人脑的感知机制。语音的感知问题属于语音的心理特征。笼统地说，语音通过空气传播到人耳，使耳鼓膜振动而转化成神经脉冲抵达人脑，从而形成感知。

人脑的左半球主管言语的接收和感知。与语言有关的中枢有言语发动中枢，听觉中枢，呼吸中枢，口、舌、喉中枢，书写中枢和阅读中枢。

语音的心理特性带有很大的社会属性。例如英文有 [p]、[b] 轻浊之分，属于不同的音位，pig（猪）与 big（大）的区别仅在于它们词首辅音的音位不同。英文中轻辅音送气与否不构成音位变化，例如 [p‘] 与 [p] 是相同音位的条件变体，即使不加区别也不影响理解。英语 speak 中的 p 在 s 后一般不送气，但即

使读成送气音，因仍属同一音位，无碍于接收和感知。但在汉语普通话中情况则不同，汉语普通话没有浊辅音 [b]，相应的轻辅音按送气与否分为 [p‘] 和 [p]，构成不同的音位，因而具有重要的辨义作用。例如“怕”与“罢”的语音区别仅在于声母是否送气。由此可见，语音的心理特性由于涉及语音的感知问题，因而受到一定语言的社会环境的制约。

9.3 实验语音学发展简况 本世纪初，词音分析主要靠一些标音符号予以记录。国际音标是国际语音协会在 1888 年公布的，后来几经修改，日趋完善和精密。图 1 所示为修改至 1979 年的国际音标方案(见下页)。

实验语音学最初也用喉镜等医疗器械研究发音动作和口型特征。30 至 40 年代采用 X 光照相等手段加强了言语生理特征的实验研究。50至 60 年代，声谱仪已经普遍应用于语音研究，实验仪器不断完善。如 X 光电路照相：动态颚位记录装置等仪器可测到语音的动态分析数据。在理论上，语音研究渐成体系，总结出许多语音规律，其中区别特征 (distinctive feature) 理论的作用尤为显著。

根据区别特征理论，区别词义的最小单位还不是音位，而是比音位更低的特征位。也就是说，音位也是可以解析的。以 [n]、[l] 为例，一些人往往区别不清，例如“男”、“兰”不分。但追溯下去，n 、l 之间还有一些共同的语音特征位，例如它们都是浊辅音和舌尖音，所不同的只是鼻音和边音的区别，而只有这个区别才真正起到辨义作用。根据雅可布逊的声谱分析法，提出了12 个对立的特征位，如元音性/非元音性，鼻音性/口音性，等等。某一音素具有对立面的前一个特征的用 “+” 号表示，具有对立面后一个特征的用“-”号表示。这种逐层区分音位的二分法尤为适用于电子计算机的语音信息处理。图 2 为英语中部分音位的区别特征矩阵表。

70 年代以来，随着电子计算机在语言研究中的应用，也给实

图 2　区别特征矩阵表

	o	a	e	u	ə	i	l	ŋ	ʃ	tʃ	k'	ʒ	dʒ	g	m	f	p	v	b	n	s	θ	t	z	ð	d	h	(*)
1.元音性/非元音性	+	+	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2.辅音性/非辅音性	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	–
3.聚集性/分散性	+	+	+	–	–	–		+	+	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–		
4.低沉性/尖峭性	+	+	–	+	+	–									+	+	+	+	+	–	–	–	–	–	–	–		
5.抑降性/平坦性	+	–		+	–																							
6.鼻音性/口音性								+	–	–	–	–	–	–	+	–	–	–	–	+	–	–	–	–	–	–		
7.紧张性/松弛性									+	+	+	–	–	–		+	+	–	–		+	+	+	–	–	–	+	–
8.延续性/突发性									+	–	–	+	–	–		+	–	+	–		+	+	–	+	+	–		
9.刺耳性/圆润性										+	–		+	–							+	–		+	–			*

验语音学增加了现代化的实验工具。有关语音的物理特性的研究不断深化，言语合成与识别的实验不断加强。实验语音学与自然语言理解、人工智能、机器翻译等领域相结合，形成了不少新的跨学科的研究和应用领域。

上面谈到，音质是语音物理特性中最为重要的一个。通过研究的深化，人们发现：不同的音质在声音的基音和陪音的关系方面有种种不同的情况，自然界的声音大部分为"复音"，象音叉发出的简单的"纯音"的情形是颇为少见的。根据傅利哀分析法，复音可分解为若干纯音，一个复波可以由若干个谐波组成。其中一个频率最低(音高最低)和振幅最大（音强最大）的谐波叫作"基频"，其余的叫作"谐频"，即陪音。陪音频率是基音频率整数倍的是乐音，否则是噪音。元音都是乐音，辅音多为噪音。一个声音的音质取决于谐波的数量、大小和与基频的组合方式。

实验语音学对共振现象作出了科学的解释。如果两个物体(如音叉)具有相同的固有频率，那么第一个物体振动时可引起另一个物体的相同振动，这种现象就是共振。利用共振原理，可利用一个共振体测得一个原振动体的频率。这种办法叫作滤波。而语音信息处理中应用的滤波器正是这种共振体。它允许以一定频率为基准的包括一定高、低频率范围的声音通过。这一范围也叫作"带宽"。

言语是由声道中空气的振动引起的。声道一头是声带，另一头是嘴唇和鼻孔，其间有口、咽、鼻形成的复杂的共振腔。语音就是由共振腔的空气振动所引起的。不同的元音发音部位不同，声道的形状不同，口形不同，共振频率也不同。一个元音的特征是由特定的能量集中区域——共振峰来体现的。

随着对语音研究的深化和计算机的普及，实验语音学的重点已转向语音信息处理的各个方面。其中主要是言语合成与言语识别。下面分别予以介绍。

9.4 言语合成 本节主要介绍用计算机合成汉语的研究，其中首

先应当提出中国社会科学院语言研究所实验语音学研究室研制的汉语普通话音节合成系统。① 这个系统已能成功地随机合成出各种声、韵、调结合的普通话音节。该系统在分析普通话音节结构的基础上，采用共振峰音节合成器，将音节划分为若干音段，输入控制各音段有关语音特性的参数，以合成程序为制导，分段合成。这些语音参数是以数据文件的形式建立在计算机外存贮器之中的。合成时调入内存贮器，经数/模转换器和低道滤波器输出。

合成是逐帧进行的。帧长的选取要适宜，一般为5毫秒。帧长过长不能很好地体现语音的连续变化，合成效果不理想，而帧长过短则又费力误时，也不合算。为达分段合成之目的，首先要对音节基本结构作深入的研究，对各个音素的发音全过程进行动态模拟。众所周知，普通话音节是由声母和韵母拼读而成的，其中声母可能包括五个发音阶段：（1）无声阶段，（2）爆破阶段，（3）摩擦及/或噪音阶段，（4）送气阶段，（5）与韵母的过渡阶段。韵母可能包括四个发音阶段（5-8接续排号）：（5）声韵母过渡阶段，（6）介音段，（7）主要元音段，（8）韵尾段。由此可见，鉴于阶段既属于声母又属于韵母的过渡阶段，音节构成阶段事实上至多为八种。

应当提出，各个阶段并非总是共现的。不同的音素情况不同。例如普通话的擦音 f、h、sh、r 等仅包括第3和第5阶段，但塞音 b、p、d、t 等都包括1至5各个阶段。

汉语普通话有六类22个声母，三类38个韵母和4种声调。由这些声、韵、调的不同组合可构成1260个不同的音节。普通话的声、韵、调分类表参见图3。

为了对普通话音节进行分析与合成，必须总结出有关声、韵、调的各种参数及组成音节时的基本连接规则，并将有关数据存入合成器的数据文件中。目前，这个在微机上建立的言语合成系统

①参见中国社会科家院语言研究所语音合成课题组《普通话音节合成系统的基础研究》。

已能根据键入的汉语拼音，从合成器中自动提取参数，按相应的规则随机合成出任何普通话音节。

图 3　普通话声、韵、调分类表

声母	塞音	b, d, g, p, t, k	6
	擦音	f, h, s, sh, x, r	6
	塞擦音	z, zh, j, c, ch q	6
	边音	l	1
	鼻音	m, n	2
	零声母	Ø	1
韵母	单韵母	a, i, u, ü, ü, ɿ, ʅ e, o	8
	复韵母	二合：ai, ei, ao, ou, ia, ie, ua, uo, üe	9
		三合：uai, uei, iao iou	4
		儿化：er	1
	鼻韵母	an, en, ang, eng, ong ian, in, iang, ing, iong uan, uen, uang, ueng üan, ün	16
声调		阴平（高平调）˥(55)　阳平（中升调）˧˥(35)　上声（降升调）˨˩˦(214)　去声（全降调）˥˩(51)	4

9.5 言语识别

9.5.1 ASR 综述　言语识别是将连续的言语声学信号转换成若干不连续的、具有一定意义的可以理解的语音片段的过程。如果无须人的直接干与即可自动完成这个从声学信号到语言信号的转换，那么，这一过程则称为自动言语识别（antomatic speech recognition 简称 ASR）。计算机言语识别是自动言语识别的一个特例。因为自动言语识别除利用计算机承担之外，还可以用其他专用设备和机器进行。当然，计算机言语识别比以往其他任何

形式的 ASR 都更加先进和灵活。

同计算机进行言语通讯实际上是从功能上模拟人与人之间的通讯渠道。当计算机用来充当言语输出装置时，它实际上是在模拟讲话人，其过程乃是将特定的符号和概念转换为连续的声学信号的智能过程。这个过程叫作言语合成或言语生成。这一问题我们在上一节已经谈到。另一方面，当计算机接收讲话人发出的言语信息时，它所涉及的是模拟人的耳→脑感知系统接收语音信息的过程。这一智能过程即是自动言语识别。计算机接收语音信息的最终目的在于使之产生一个人们所需要的反应或回答。一个成功的言语识别系统不仅能将接收的语音信息转换成一个个的相应的词语。而且，为了归纳语音信息的意义，计算机还必须确定句法结构和语法关系。显然，这些都涉及人工智能的研究领域。

计算机的反应回答机制可以靠预先确定的语言形式，也可以靠随机合成。显然，后一种较为复杂。人机对话系统是自动言语识别与合成的结合体。回答方式同言语识别器的设计有密切关系，而言语识别器的设计又决定了整个人机言语对话系统的质量。

有关言语识别的研究存在着几种不同的考查角度，基本上是从言语链环的不同方面考虑，分别探讨言语识别问题的。

言语产生的观点旨在以人类发音系统产生语音的方式为基础，分析这些语音信息究竟承载了哪些参数。

言语的声学观点认为言语是一种声学信号，所以应当采用在物理分析中普遍有效的信号分析技术，如频谱分析法等。

言语感受的观点则试图根据人类感受语音的典型方式建立一些有关的参数。例如使计算机模仿人的感受接收过程，从听觉感受的角度提取一系列的语音参数。

言语识别是一个跨学科的课题。它涉及语言学、语音学、生理学、声学、心理学的综合研究。设计出可以识别人类语音的机器，进而建立人机语言对话系统，是言语识别和言语信息处理的目的。

9.5.2 言语识别的若干方法 首先介绍一下利用区别特征进行言

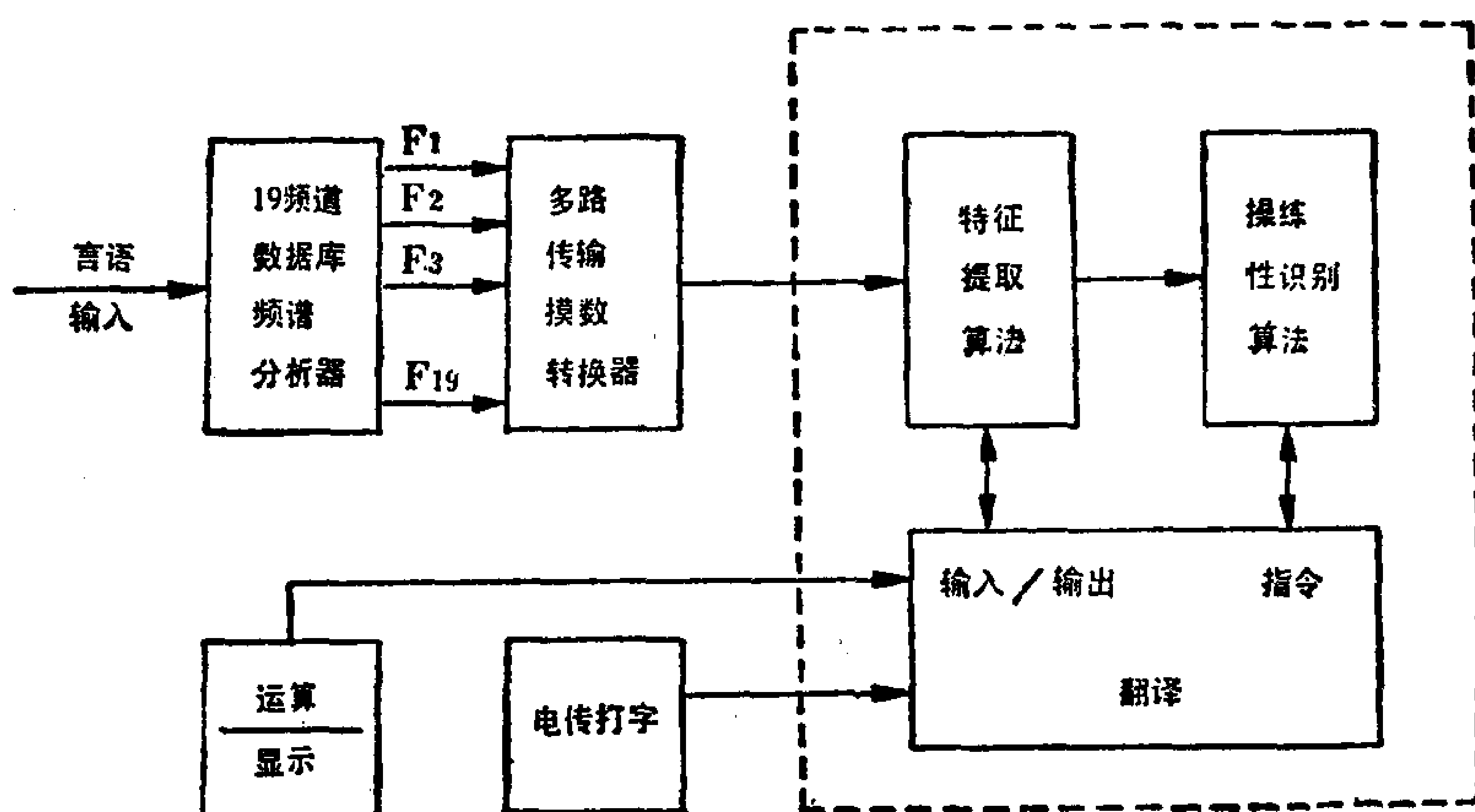
语识别的方法。涉及言语识别的区别特征是从基本声学参数中推导出来的。重要的声学参数有：共振峰频率、共振峰位置、频谱形式，以及是否产生噪音或无声等等。根据这些特征，在同语言学相关的区别特征的层次进行模式匹配。计算机程序根据5种声学特征来确定区别特征，而这些区别特征也用来对词语的音位段进行分类。

关于连续话语的识别，已有专用软件以解决连续话语识别器的一些关键性的设计问题。例如如何把言语切分成若干还连续的较小的识别单位，以及如何把自动机理论应用于设计一种用于人机言语对话的专用语言。识别器以音节为单位进行整体识别。而不必拘泥于音位分析，并在连续话语的语言分析方面进行言语的超音质的监控。实现言语的连续识别以及转换为文字符号予以输出是语音打字机研制的目的。

麻省理工学院另辟蹊径，研制了一种针对54个词语的识别器(GOLD 系统)。这个识别器采用16频道的频谱分析器、音高提取器以及元音探测器作为它的硬件组成，通过放大器和模→数转换器将信号输入计算机。计算机程序将输入的词语进行切分和分析。当试讲人朗读词表内的词语时，计算机对切分音段的各种测试结果进行分析和归纳。在此基础上建立了一种决策算法，以表内的词语作为数据，通过加工分析得到识别结果。一些专用程序根据词语前后的无声段用来监别词语的开头和结尾。词语内部的切分则根据言语信号的音强、浊音的起止点、频谱形态的变化、共振峰及其位置等因素而建立其算法的。根据对每个词的切分段进行的分析，与预先选定的范畴加以比较，即可将各个词语区分开来。这一算法经十个试讲人的检验，识别正确效达86%。

采用 GOLD 系统的技术手段，并吸取以往言语及图像识别的经验，又研制了一种有限言语识别器，这就是 LISPER 系统。在这个系统中采用样品数据库向计算机软件提供识别依据。这一系统包括特征提取和操练性识别算法(参见图4)。LISPER

图 4 LISPER 系统示意图



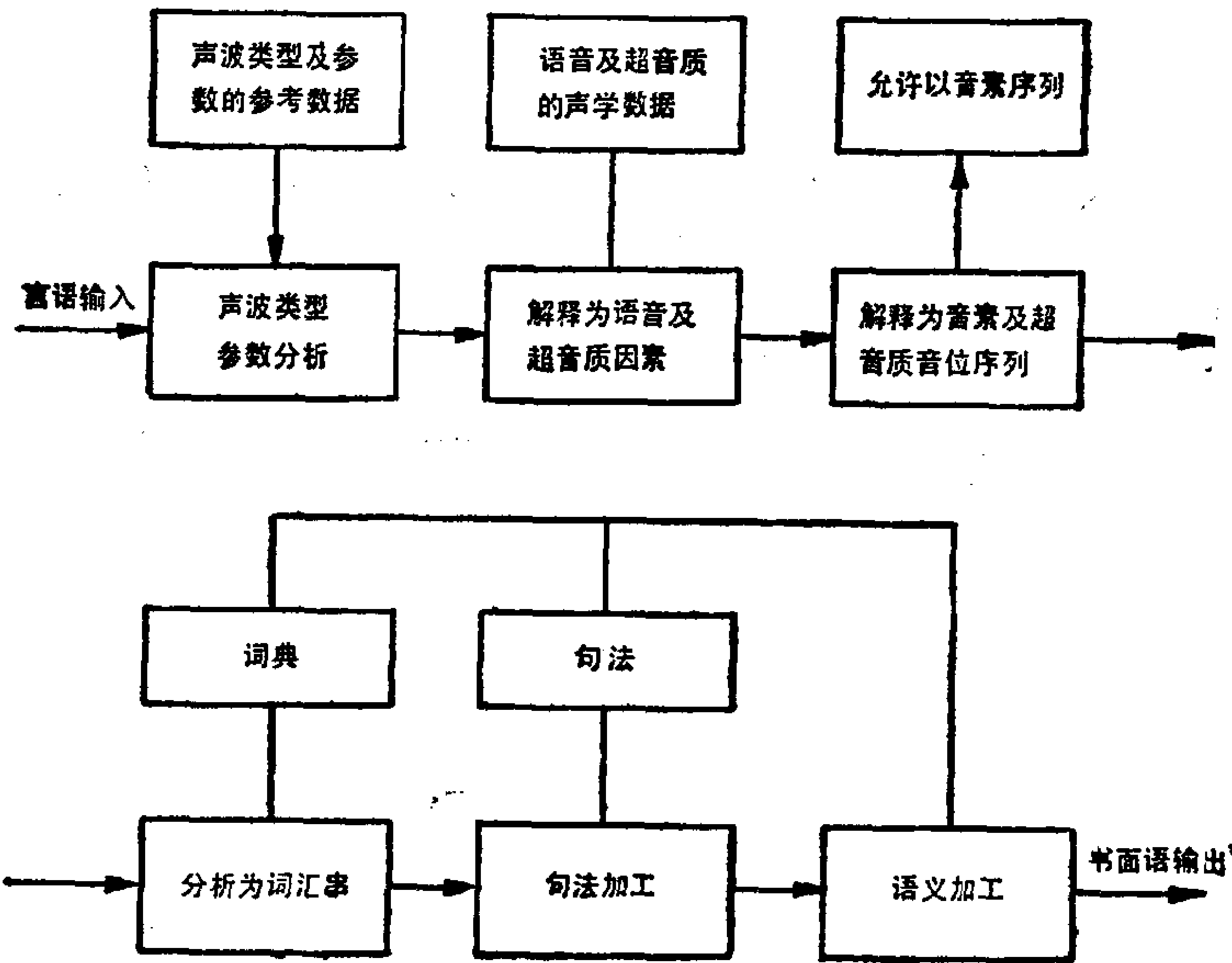
系统采用学习机的方法，只要试讲人将每个词重复地读几遍，通过指令翻译器传输给机器，即可判断每一词条与哪个特征模式相关。这些模式都是事先经操练予以存贮的文件记录。一旦同输入的话语匹配成功，即实现了识别。特征提取软件可根据频谱分布确定有关的语音特征。

9.5.3 言语识别的若干理论问题　在60年代初，言语识别的理论已初步建立。根据音韵学和区别特征的理论提出了几种选择ASR技术的原则。有人探讨如何运用声学及语言学参数将言语的声学符号转换为不连续的音素代码。例如 Peterson 在60年代例举了与言语自动识别有关的声学特征，并强调指出 ASR 是追溯语言学的代码。他主张动态言语分析法。这种分析法的特点是，首先采用声学线索进行识别，对构成话语的音素进行猜测。鉴于单凭声学信息不足以精确地进行识别，剩下的歧义问题可以借助存贮的语言学信息得以解决。假如 /m/、/n/ 不易从声学上识别清楚，在和元音 ü[y] 相拼的时候，还可以根据语言学信息进行分析。在特定的语言环境中，可以用排除法进行决择(例如在汉语普

通话中 nü 是可接受的音节，而 mü 则是不可接受的)。由此可见，如果在机器中存贮有音素搭配组合规律的信息，就容易解决语音识别歧义的排列问题了。这一观点可从图 5 (Peterson 自动言语识别总体示意图) 中得其基本宗旨。这种声学及语言学信息的综合利用法首先是声学分析，之后再用语言规则(例如音段模式信息)作为补充手段进行核实或再判断。

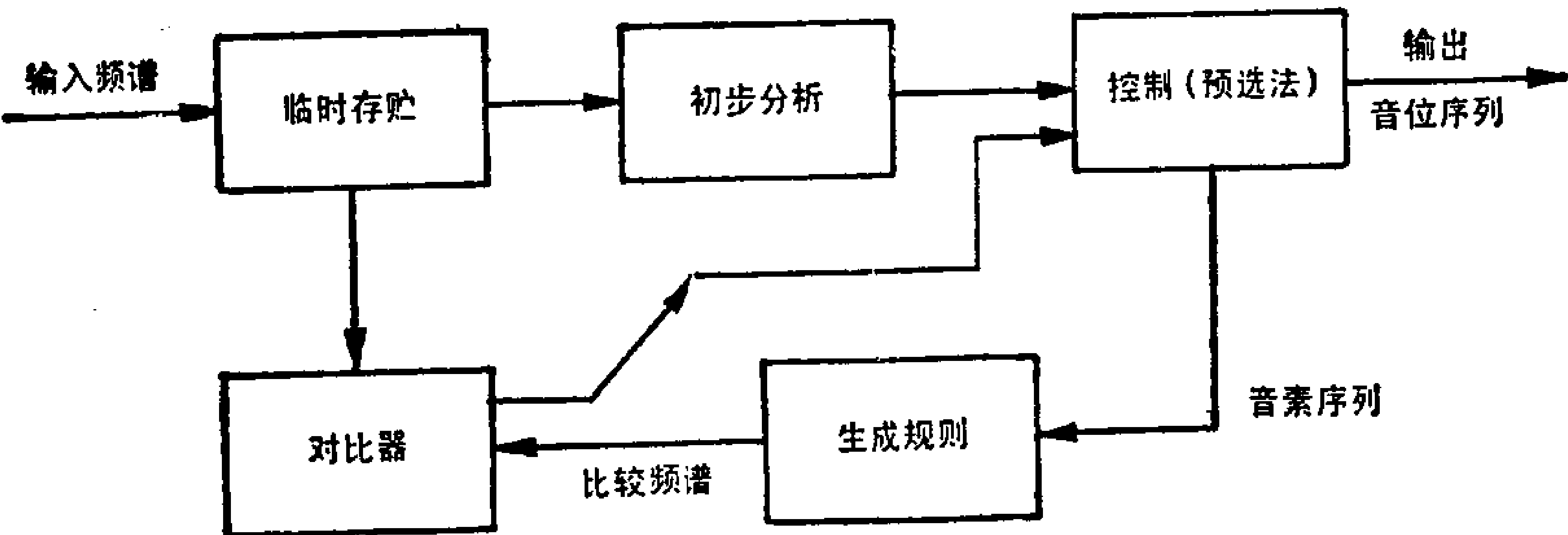
另外一种合成法分析模式则具有一系列的生成规则。这些规则能够合成那些可同输入信号进行比较，直至实现理想匹配的诸多信号。这里的合成器取代了一般存贮模式的词典，并能够从较少的规则出发生成众多的模式，以便同输入信号进行比较。合成器具有选择控制，决定首先合成哪些信号，通过接收的信号进行初步分析，对其音质及特征进行预测，这样就可以将与输入信号相

图5　Peterson自动言语识别总体示意图



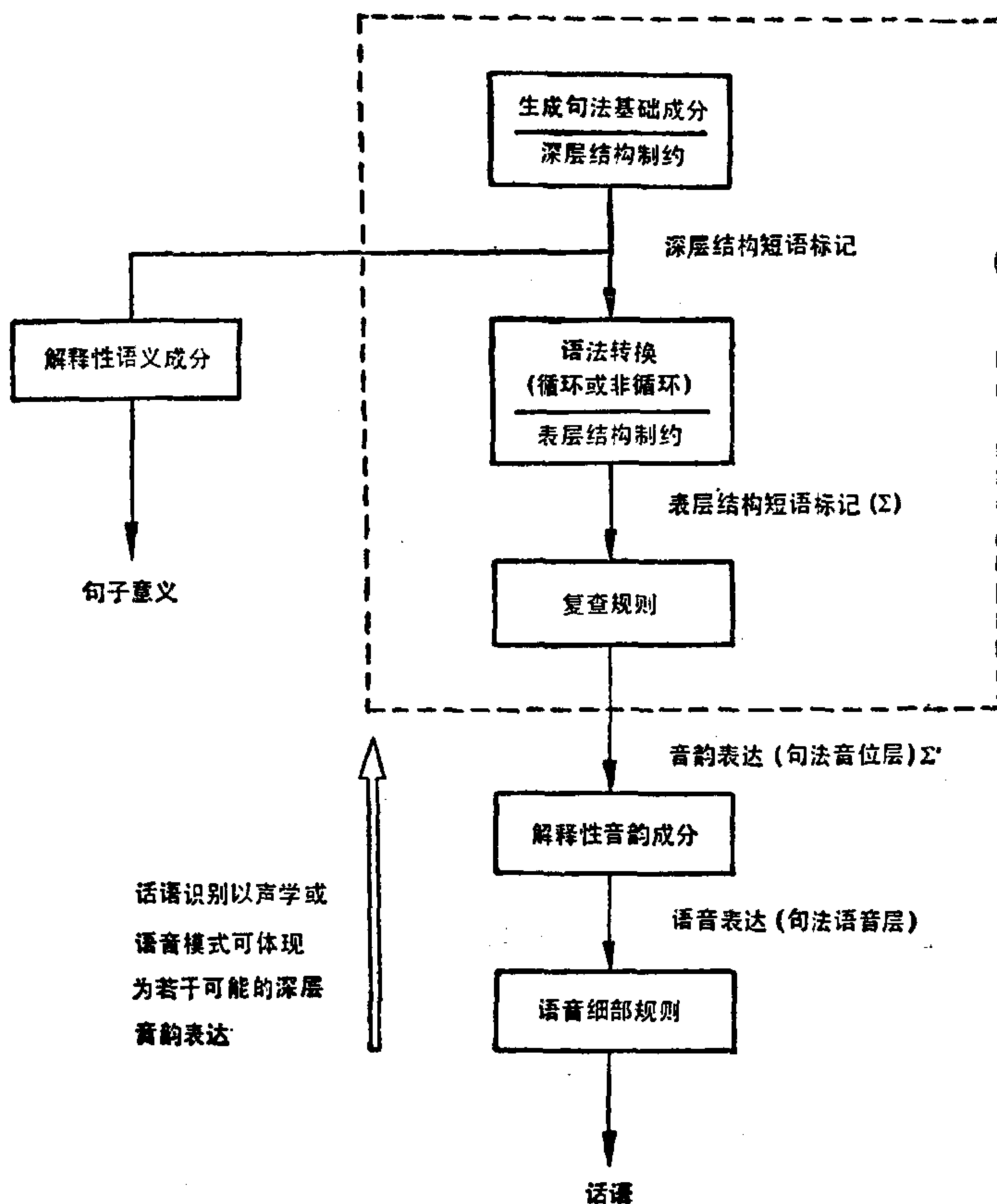
对应的可能的音素序列限制在一定的范围。每一组送入生成规则的可能的音素序列都产生一个信号(参见图6合成法分析模式)。这个信号可能是频谱，也可能是与输入形式相匹配的其他层次的信号。对比器则用来判定输入信号是否与生成的模式相匹配。如果不匹配的话，则告诉控制器究竟有哪些不同。根据初步分析的结果和对比器中的错误信息及分析数据，控制器即可决定下一步应当尝试哪一个音素序列。当实现了正确匹配的时候，对比器则告诉控制器输出，用以合成正确信号的音素序列。

图6 合成法分析模式 (Halle & Stevans, 1964)



根据 Chomsky 的观点，将一个被考查的声音产生过程视为某种特定的语音序列，即是确定其句法结构。根据生成语言学的理论，音素不能离开句法而独立存在。根据言语分析的研究，取代独立音位层的应当是按照生成理论形式的系统音位层。这一系统音位层也叫语素音位层，也就是不仅将言语视为语音信号，也视为句法和语素音位信息，这就涉及到语素、词语界限、表层句法结构及语法构形成分等等。图7是转换语法示意图，它描述了言语产生的语言学过程。从中我们可以看到系统音位层体现了转换生成句法的结果。系统音位学还包括一整套有序的解释性的音位规则，用来给句法结构赋与语音形式。这种解释性的音位规则在语法和语音之间架起了一座桥梁。

图7 转换语法示意图



9.6 结语 实验语音学的研究在整个语言学体系中不是孤立的。为了更有效地进行语音的信息处理，有必要涉及到句法和语义的研究。因此，在篇章结构中研究言语的识别与合成，进而实现理想的人机语音对话智能通讯系统和语音翻译机，应当成为语言工作者和信号工程专家为之奋斗的重要目标。

10. 中文信息处理

10.1 中文信息处理

10.1.1 弄清几个概念 中文是“中国的语言文字，特指汉族的语言文字”(见《现代汉语词典》)。这里的“中国的语言文字”不等于中国各民族的语言文字。它的含义是：中文是中国的官方语言，换句话说，汉族的语言文字是中国的官方语言。例如，当我们说中文是联合国的工作语言之一，这个中文就指的是中国的官方语言，或汉族的语言文字。

什么是“信息”呢？信息是一个极为广泛的概念，定义有好多种，但通常的理解是，凡是能通过视觉、听觉、嗅觉、味觉、触觉等器官或仪器获取，并且有一定交际功能的东西均可视为信息。信息既具有一定的形式，又具有一定的内容或意义，是一个统一体。但是，我们所要处理的信息只包括通过视觉和听觉得到的信息，最多还包括通过触觉得到的东西(盲文)。从这里不难看出，中文信息不仅包括汉字信息，而且包括汉语信息。

所谓处理，就是用电子计算机对信息进行各种加工，概括地说，不外乎图象信息和语音信息的识别、模拟、分析和转换。

因此，我们可以说，中文信息处理指的是利用计算机对汉语书面形式和口语形式这两种信息进行各种加工，而加工的结果是形成各种信息处理系统，如中文情报检索系统、汉语语音识别系统、英汉机器翻译系统，等等。

但是，对中文信息处理这个术语常有两种误解，一是把它的

概念缩小，等同于汉字信息处理；另一是把它的概念扩大，理解为中国各民族语言文字的信息处理。

汉字信息处理系统只是中文信息处理系统的一部分，也是非常关键的一部分。具体地说，它是一个工具，没有这样的工具，中文信息处理系统建立不起来，因为汉字是当前我们社会上通行的文字。但是，汉字信息处理系统研制成功，一般只是解决了汉字的编码、输入、存储、编辑和输出问题。至于加工或处理什么，如何加工，那是中文信息处理的内容。中文信息处理系统（纯口语的系统和拼音文字的系统除外）除了以汉字信息处理系统作为自己的必备部件外，本身还带有为不同目的服务的各种应用软件。它的另一特点是一般都以词作为加工的基本单位（汉字频率统计可说是个例外）。比如说，要把"我想学习英语"这句话译成外语，首先要把词切分出来，即切分出"我 想 学习 英语"这四个词，才能进行翻译。"词"是翻译的基本单位。

中文(汉语)是许多自然语言中的一种，因此，中文信息处理又是自然语言处理的一部分。我国有几十种少数民族，他们的语言文字信息处理也是自然语言处理的一部分。

自然语言处理又是人工智能(即用计算机模拟人的智能活动)这个极为广阔的领域的主要内容。信息有多种，其中语言信息是人类社会中最主要的信息。任何一种物质或任何一种精神存在，只有转换成语言信息，才能为人所认识和理解。前一阶段，人工智能的主攻方向没在语言方向，这是同计算机的水平不够高有关的。第五代计算机要解决这个问题了。从这个角度上说，自然语言处理尚处于方兴未艾阶段。

弄清汉字信息处理、中文信息处理、自然语言处理以及人工智能这几个概念，非常必要，这不仅能帮助我们了解它们各自的范围和它们之间的关系，而且能帮助我们理解过去，观察现状和展望未来。

10.1.2 中文信息处理研究的范围 尽管中文信息的内容极为广

泛，但(计算)语言学工作者所研究的只是中文信息处理的一般理论和方法，以及与语文信息加工有关的基础性工作。语言工作者所研究的问题，与诸如考古学上的鉴定、经济学上的市场预测等专业内容无关，但确是各专业都离不开的一些最基本的东西，例如，汉字如何编码才能输入计算机，如何分词，哪些是常用词，如何进行语音识别与合成，等等。

目前已经开展的中文信息处理研究项目有如下一些：(1)汉字信息处理，(2)言语统计，(3)索引、词表和词典的编制，(4)情报检索，(5)计算机辅助教学(程序教学)，(6)自然语言理解(人机对话)，(7)机器翻译，(8)语音识别与合成，(9)方言研究，(10)风格学研究，(11)术语数据库。

随着研究手段的改善和研究工作的深入，还将有更多更新的项目涌现。社会不断进步，新技术革命不断深入，中文信息处理的范围也将相应地扩大，以满足社会的需要和迎接新技术革命向我们提出的各种挑战。

10.1.3 中文信息处理和“传统”语言学的关系 中文信息处理是语言学同计算机科学、自动化技术等科学部门相结合的产物，是一个新兴的边缘学科。它的发展有赖于这几方面专家的共同努力。但它与语言学的关系更为密切。传统语言学解决的问题是针对人的，而它所要解决的却是“人—机—人”的问题。传统语言学是它的基础，对它的发展有直接影响。例如，普通话的推广将为语音识别创造极为有利的条件。又如，汉语词类如得到合理划分，将大大有利于汉语理解和汉外机器翻译的分析工作。再如，号称信息处理的“拦路虎”的词，如得到科学规定，将对中文信息处理各个方面的工作产生巨大作用。反过来说，中文信息处理对传统语言学也不是无所作为的，它能在很多方面对传统语言学加以检验、订正和补充。例如，利用计算机对语言文字各种要素(元音、辅音、声调、声母、韵母、音节、部件、字、词、词组、句)的统计，可以提供量的描述并据此给出质的评价，从而丰富和发展传统语言学。

10.1.4 中文信息处理的意义　目前，人们都在谈论世界正进入信息化社会。从语言学角度来说，信息化社会最主要的特征就是利用电子计算机等先进工具对语言文字信息进行各种处理，目的是建立现代化语言信息系统，使语言文字得到最佳利用，使凝聚在语言文字中的知识信息发挥最大效能。

大家知道，语言文字的应用水平直接反映一个民族或一个社会的发展程度。技术先进的民族已使用电子邮政，他们的办公室已在很大程度上自动化，他们还在航天飞机上举行了记者招待会。但是，有的民族，直到今天还没有文字，这就是语言文字应用水平的差别。

我国是一个古老的国家，在历史上我们的语言文字曾达到高度的应用水平，创造了丰富的文化遗产。但是，由于各种原因，在采用新技术提高语言应用水平方面比国外落后了一大截。为了迅速改变落后面貌和促进四化建设的进程，有必要在我国尽快建成现代化语言信息系统。这个系统是一个极其复杂的巨系统，其中包括汉字信息和语音信息两个大系统，而每个大系统之下又有一些分系统和小系统，另外，在不同系统之间还组合成一些混合系统。这个系统的建成，将标志人工智能的高度发展，同时也标志着信息化社会的到来。

10.2 汉字编码输入技术

10.2.1 汉字编码是汉字进入计算机的关键　中文信息处理研究中最关键的一个问题就是汉字信息处理系统的研制。汉字信息处理一般包括：编码、输入、存储、编辑、输出和传输。汉字的编码输入是关键的关键。不解决这个问题，汉字就不能进入计算机，图书情报工作自动化、印刷排版现代化、办公室事务自动化等都将落为空谈。

汉字编码指的是如何给汉字规定一种便于计算机识别的代码。至今，这样的编码方案已有四百多种，其中上机通过实验的和已被采用作为输入方式的也有数十种之多。归纳起来不外五种

类型：(1)整字输入，(2)字形分解，(3)字形为主、字音为辅，(4)全拼音输入，(5)拼音为主、字形为辅。

10.2.2 五种类型编码法简介　(1)整字输入法：前一阶段，一般是将三、四千个常用汉字排列在一个具有四、五百个键位的大键盘上，近来，大多是将这些汉字按 xy 坐标排列在一张字表上。后一种通常叫"字表法"或"笔触字表法"。电笔点触到哪个汉字，那个字便随即输入。比如，x 25 行和 y 90 列交叉的字为"国"，当电笔点到字表上的"国"字时，机器自动输入该字的编码 2590。键盘上或字表中的排列方法，有按部首的，有按音序的，也有按字义联想的。不常用的字作为盘外字或表外字，另行编码处理。

(2)字形分解法：将汉字的形体分解成笔画或部件，按一定顺序送进机器。笔画方案中一般归纳出八种笔画：横(一)、竖(丨)、撇(丿)、点(丶)、折(乛)、弯(乚)、叉(十)、方(口)。部件方案中，一般归纳出一、二百个部件，有的还多些。大家知道，普通打字机键盘上只有 42 个键（包括数字和标点），装不下这么多部件。为此，有人设计一种中键盘，也有人利用部件形体上的相似点或出现概率的不同，而把一百多个部件分布在 26 个字母键上。

(3)字形为主、字音为辅的编码法：这种编码法与字形分解法的不同点在于还要利用某些字音信息。如有的方案为了简化编码规则，缩短码长，在字形码上附加字音码；有的方案为了采用标准英文电传机，将分解归纳出来的字素通过关系字的读音转化为拉丁字母。

(4)全拼音输入法：绝大多数是以现行的《汉语拼音方案》为基础进行设计。关键问题是区分同音字，因而有的方案提出"以词定字"的方法，还有的方案提出"拼音汉字转换法"，即"汉语拼音输入——机内软件变换(实为查机器词表)——汉字输出系统"。

(5)拼音为主、字形为辅的编码法：一般在拼音码前面或后面添加字形码。拼音码有用现行《汉语拼音方案》或稍加简化的，还有为了缩短码长而把声母和韵母都用单字母或单字键表示的

“双拼方案”或“双打方案”。如 F 键既表声母 F，又表韵母 ang，连击两下，便是 fang “方”字。区分同音字的字形码也多种多样。大部分采用偏旁部首的信息，还有采用起末笔的，采用语义类别的。

10.2.3 综合评述　上述各种编码法，各有所长，也各有所短。例如，字表法的特点是一字一格（键）无重码，直观性好，操作简单。缺点是需特制键盘，熟练时间较长。字形分解法的好处是按形取码，不涉及字音，因而不认识的字（包括生僻字、古字）也同样可以编码输入。但汉字形体结构非常复杂，写法也有许多差异，分解标准不易统一，因而不少方案规则较多。为了提高速度，有些方案把特别常用的字和难分解的字作整字输入，另外还采用了一定数量的词汇码。拼音输入法的优点是操作简捷，可以“盲打”，不受汉字简化、字形改变的影响，符合拼音化方向。缺点是不认识的字无法输入；另外，如果不加字形码或不用以词定字法或显示选择法，同音字较难处理。为了提高速度，一般采用“双拼双打”键盘。同时，还采用了不少类似外语缩略语形式的词汇码。如 DGWZ “德高望重”。“拼音—汉字转换法”对一些用户使用方便，且有利于进一步的中文信息处理。如果考虑到汉语拼音的推广应用，会有更大的发展前途。因此，美国 XEROX 公司、日本 Toshiba 公司、英国 Sindex 公司都在朝这方面努力。

10.2.4 采用双轨制好　汉语拼音推广应用，并过渡到汉字和汉语拼音文字并存并用，这是一种双轨制，也是一个方向。中文信息处理领域中音码和形码的并存并用同样是一种双轨制，也同样是一个方向。因此，不少人认为，采用双轨好，理由是：

(1)对掌握普通话的人来说，使用音码比形码方便，而且快一些。形码虽然较慢，但能输入任何汉字（包括古字）。采用双轨，相得益彰。操作员认识的字可按音输入，不认识的字按形输入；会普通话的人按音输入，方音重的人可按形输入。

(2)对于用字量少的单位，按音输入无问题，但对用字量大

的单位来说，按音输入就不如按形输入，因为一般人只能念出一部分汉字的音。

(3)按形输入(尤其是整体输入）对于中文信息处理的某类工作，例如统计汉字非常适合，但是对于其他工作，例如统计汉语语音(元音、辅音、声、韵、调……)，则无能为力。按音输入则正相反。由此可见，双轨正好是相辅相成。

(4)形码还有一些优点，如有的形码可以照顾多种汉字(如日本的、南朝鲜的)。音码还有一个大优点就是分词连写，便于信息处理。汉字文献无词的界限，而中文信息处理的基本单位不是字而是词，这是现代汉语的特点所规定的。音码和词是一对孪生子，正好满足这一要求。

(5)适当的双轨方案不会增加设备上的麻烦。如不考虑整字输入，一般可使用现有小键盘。

10.2.5 汉字编码研究的新发展 除了单轨向双轨发展之外，还有下列趋势：

(1)混合式编码法：这是相互学习，取长补短的结果。例如，笔触字表法中除整体字之外，又增加了一些部件或字元，这样，不但解决了表外字问题，而且大大提高了它的功能，搞得好的话，可以囊括字形分解法的优点。又如笔画方案为了提高速度也增加一些部件或整字。

(2)充分利用简码和词汇码，以提高输入速度。字频和词频已成了字和词的形、音、义之外的第四特征，这不是夸张的说法。大家知道，“的”“一”“是”“在”“了”“不”“和”“有”“大”“这”等十个字在两千多万字的文献中出现二百四十多万次，占 11.40%，其中“的”一个字就占总数的4%弱。编码工作者充分利用了这一特征，设计了单字母和双字母的简码，如以 d 代表“的”，以 b 代表“不”，等等。

词汇码是又一得力的手段。有一种形码方案的词汇码是根据每个字的部件规定的，如“汉字编码”的词汇码是43、45、55、13

(シウ系石)。另一种编码方案的词汇码是利用计算机辅导方式输入的。例如，当“中”字输入后，一按语词键，屏幕上便显示出“中国”“中型”“中性”“中华”等双音词；选择“中国”后，如再按一下语词键，便可显示“(中国)话”“(中国)人民”“(中国)共产党”“(中国)工农红军”等多音词。音码方案的词汇码实际上为词组码，如ZRG“中华人民共和国”，ZZXY“中国中文信息研究会”。词汇码不但能提高速度，而且也能区别同码。但是，如果用得太多，就会产生重码。因此，有必要划分通用词汇码和专业词汇码，以减少重码。

(3)充分发挥“电脑”的作用，尽量减少“人脑”的负担。上述计算机辅导输入法就是一例。还有的方案不断以开天窗方式向操作员提供选择的范围。这样，操作员不必再记忆大量的编码规则。

10.2.6 编码工作中的标准化和定型化　编码方案繁多，如果没有一个标准加以统帅，那就要发生混乱。1981年，国家公布了《信息交换用汉字编码字符集基本集》(简称汉字标准交换码。汉字共分两级，一级3,775字，二级3,008字，共6,783字。这种汉字标准交换码是计算机的内部码。有了它，各种输入输出设备的设计就有了统一的根据，各种系统之间的信息交换就有了共同一致性，信息资源的共享也便有了保证。目前，正在制定《信息交换用汉字编码字符集辅助集》，以便满足少数用字量超过基本集的用户和台湾、香港等地的需要。

编码方案的定型化，即一般所说的优选工作，也非常重要。汉字编码的“战国时期”应该结束了，否则，对计算机的普及应用将带来不良影响。当然，定型或选优并不意味着只定一种或只选一种，要照顾到多种用户的需要。关于选优，曾经提出多种评定指标，一般包括字码无二义性、操作方便易学、输入和处理效率高、存储节省、传输可靠、设备经济实用。但是，有一项重要指标没被重视，这就是组词能力。如果输进的码是一堆没有词的界限的汉字，这将给下一步的中文信息处理带来很多麻烦。求助机

器进行自动切分，不仅要占用机器大量宝贵的时间和空间，而且目前也得不到切实的保证。从这一点来说，音码占居上风。

10.3 汉语词库问题 目前，研究中文信息处理的专家们正在酝酿建立词库，这是一件大好事。这说明，中文信息处理已发展到一个新阶段，即由字处理过渡到词处理阶段。换句话说，人们已认识到，词是中文信息处理系统中的基本加工单位，为了建立各种信息系统，必须树立词的概念，解决好词的问题。

10.3.1 词库　词库实质上是一个计算机化的词表，为以词加工提供规范。词库可分基本词库和辅助词库两种。前者为各行各业的用户服务，后者为特殊用户提供补充。

词库的内容应该是词和一部分词组词（即带短横的词，如Wànshì dàjí，“万事大吉”），另外还包括少量的构形成分（们，了，着，过等）。因此，词库的设计应摆脱汉字的干扰，而应以汉语拼音正词法为基础。

10.3.2 建立词库

10.3.2.1 为现代汉语词汇制定计算机化的映象　大家知道，中文（汉语）发展到今天，已由单音节语变为多音节语。现在的人已不象古人那样说话写文章。例如，荀子《天论》中的第一句话“天/行/有/常/，不/为/尧/存/，不/为/桀/亡”，用现在的话来说就是“自然界/运动/有/一定的/规律/，不/因为/有了/明君/唐尧/才/存在/，不/因为/有了/暴君/夏桀/就/消亡/”。上面斜线分开的部分都是语言中一个个独立的要素（词），因此不难看出古今的差别：古代汉语中一般是一个字（音节）等于一个词，而现代汉语中，词由一个或多个字（音节）组成，并以后者为主。我们建立词库是对现代汉语词汇进行科学描述，并给它制定计算机化的映象。

10.3.2.2 为建立各种信息处理系统铺平道路　我们搞中文计算机（或称中文信息机）的目的不仅是让它能输入输出汉字，起中文打字机的作用，而且更主要的是让它能加工中文信息，建立各种

系统。然而，要做到这一点，必须以词为基础，因为词是最基本的载信单元。例如，我们要在检索系统中把“中国应该大力发展高科技”这样一篇文献检索出来，如果不以“中国”和“高科技”为叙词(关键词)是很难查到的。同样,让机器把这句话翻译成英文,也必须分出“中国”、“应该”、“大力”、“发展”、“高科技”，才能得到 China should develop high-tech energetically. 由此可见，词库是建立各种信息系统的必要手段。

10.3.2.3 为人和机器切分词树立标准　词库既然是为以词加工提供规范,它就是一种标准。作为标准，它一方面为人提供方便,另一方面也给人以约束，要求人们遵循它。然而,什么是一个词,怎样按词来写文章，这对于长期使用汉字的人来说，不是一件容易的事。因此，更需要有这样一个规范。否则,人们将无所适从,各行其是。退一步说，即使分词的工作将来由机器负担，那也要有一个标准底表作为分词的依据。因此，不管是由人分，或由机器分，总得要有一个词库作可靠的基础。

顺便谈一下，现在人们常把编码输入比作中文信息系统的瓶颈问题。这话有一定的道理，因为目前还没有一个非常令人满意的方案。但是相比之下，分词连写是一个更大的瓶颈问题，因为即使编码输入这个瓶颈问题得到解决，即使汉字自动识别和口语输入取得成功，分词问题依然存在。汉字自动识别成功之后,“人工分词”一说不复存在，非得配以机器自动分词不可。至于口语分词,那完全是另一套技术，难度更大,主要依靠重音、语调、停顿和言语概率等特征。总之，不管是书面语，还是口语，总得把“词”求出来，才能谈下一步的处理，因为“词”是现代汉语中最基本的载信单元。

10.3.3 建立词库的方式

10.3.3.1 关于建库方式和需要解决的问题　我国在建立词库方面已有一定基础，同时也有不少经验和教训可以总结。北京语言学院是开展词频统计最早的一个单位，开始阶段没有利用电子计

算机，后来在中国社会科学院语言研究所等单位的协助下将手工统计的资料进行了计算机处理。因此，他们的词库可以说是通过人工分词建立起来的。北京师范大学的词库是在统计语料过程中自动生成的，不过这些语料事先已经过人工分词，因而他们的词库可以说是半自动化的方式建成的。北京航空学院等单位进行的词频统计是目前来说规模最大的。他们的词库是在固定底表基础上通过机器自动切分生成的，因而可以说是通过全自动方式建成的一个词库。

方式不同，各有千秋。前两种词库是根据具体语料并由人工分词建成的，因而能产生真实反映统计对象的全部语词。当然，如果人工分词一关掌握不好，也会产生一些不一致现象。后一种词库与底表的关系很大，如底表制定得不够合理而又无补充措施，则不能完全反映统计对象的真实状况。但是，它可以避免人工分词中的一些不一致现象。

究竟怎样建立目前正在筹划的基本词库呢？现在已有多种词库可以作为参考，这是非常有利的条件。但是，拿这些作为基本词库的依据，似乎还不够，还有进一步做工作的必要。我们认为，在目前情况下，比较理想的方式是采用动态底表法，具体地说，就是精心选定底表，让机器进行自动切分，同时不断回收新词或机器不能切分的词，进行人工干预，从而使底表不断得到补充。

所谓精心选定底表，是指不能全盘照抄现行的词典，因为有些词典中的词项是术语不是词（如“矛盾的特殊性”），甚至是事件的名称（如“第一次国内革命战争”）显然，我们不能把这些收入词库。

利用机器进行词的切分，有利也有弊。它的最大优点是一致性强，缺点是不能“随机应变”，不能排除误差。目前使用的方法有顺向匹配、逆向匹配、双面匹配及其他改进型的匹配法。顺向匹配能解决与后字交叉的歧义问题（如：今天真热（可分为：今天，天真）），逆向匹配能解决与前字交叉的歧义问题（如：我们

从容易的做起(可分为：从容，容易，)。据航空学院的同志们证明，逆向匹配比顺向匹配误差少。双向匹配能提高精度。但速度大降。即使用双向匹配，有一些歧义现象也还解决不了。例如，人们常举的例子“结合成分子时”(可分为：结合，合成，成分，分子，子时)，怎么切分也会产生误差。另外，有些歧义现象，如“一个半劳动力就行了”(究竟是 yige-ban laodongli, 还是 yige ban-laodongli)，“乒乓球拍卖完了”(究竟是 pingpong qiupai mài wanle, 还是 ping pong qiu paimài wanle)，还得靠句子以外的上下文来确定。有的歧义更为复杂，例如，机器碰上“美国会通过生产一种新式武器”这个短句，就很难办。前三个字如何切分，就发生了问题，是“美/国会”，还是“美国/会”？两种切分语法上都成立，而且意思也讲得通。如果这是报纸上的一个标题，作第一种切分是正确的。但如果后面还有下文，而且是“来加强自己的实力”，则第二种切分是正确的。如果下文是“并批准……”，则仍应该是第一种切分。请看，这要绕多少弯子！机器要有足够的智能才能分辨清楚①。由此可见，机器自动切分还需继续改进，采取有力的纠错措施，以提高切分精度和速度。

采取动态底表法这种方式建库，耗费机时较多。如果条件不足，也还可以采用人工分词和机器自动生成词库的方式。

10.3.3.2 关于词的确定和词库的规模及结构　建库方式是一个重要问题，关系到建库的多快好省问题，不能不考虑。但是，最最重要的问题还是如何确定词的问题，而与词的确定有关的则是词库的容量问题。

什么是一个词，目前在国际上也还没有一个为大家都能接受

①如果提供这份材料的人原来就根据自已的意图分开来写（美＋国会／美国＋会），那该多省事。由此想到音码或拼音转汉字输入法的好处，它们都是以词为单位进行输入的。从系统的观点考虑，这是合算的。为此，笔者曾主张评定编码方案时，应把是否便于下一步处理(即是否便于以词加工)作为选优的标准之一。〔参见《中文信息处理技术发展现状与展望》126页〕。

的定义[1]，在中国语言学中可以说更是一个棘手的问题。由于汉字的长期影响，因而许多人心目中还没有建立起词的概念，以致时常字词不分。据吕叔湘先生考证，“首先提出‘词’作为现代语言学意义的术语，并且跟‘字’加以区别的是章士剑的《中等国文典》(1907)”[2]。因为汉语形态不发达，所以给汉语词下定义更是难事。比较通行的一种说法是：词是语言中具有固定语音形式并能独立运用的最小的结构语义单位，人们用它组成长短不等的句子相互进行交际。

尽管在使用拼音文字的语言中，从理论上对词下一个什么定义还有分歧，但实践中问题不大，两个空格（space）之间的一组字母就是词。然而，汉字文章中词与词之间没有这种标志，因而给中文信息处理造成了很大麻烦。

为了解决以词为单位进行信息加工的问题（如词频统计、人机对话等），目前一般是由人先把词分立出来，送入计算机。从已发表的一些统计材料可以看出：统计同一性质的材料而得的数据，比重各不相同，甚至悬殊很大（即使统计同一本书，也会有不同结果）。原因就是缺少词的统一标准。一般来说，多音节词划分得多，词条数目必然增加，反之，单音节词划分得多，词条数目就必然减少。问题的复杂性还在于：这样划分的结果既不符合一般的理论认识，也不便于人们的实践。更糟糕的是，如果处理不当，将会直接影响我们即将建立的词库的质量和效能。

关于词库的规模和结构，一般来说，相当于GB2312《信息交换用汉字字符集基本集》（下面简称《基本集》）功能的词库，恐怕起码要有五万词，而且其分布应该是单音节词0.05%左右，双音节词75%左右，三音节词14%左右，四音节以上的词6%左

①请参看刘涌泉《语言学现代化和计算机》中第十八节“词”，武汉大学出版社，1986。

②吕叔湘：《汉语语法论文集》（增订本）中的“汉语里‘词’的问题概述”，商务印书馆，1984。

右。为什么这样考虑呢？主要理由是：

（1）从覆盖率考虑。一般来说，2500—3000个汉字，覆盖率可达99%（文改会1985年统计，2851汉字已达99%），达到99.99%覆盖率，需5018个汉字（据新华印刷厂统计）。《基本集》收汉字6763个，完全能满足一般应用的需要。这个量定得十分合适。笔者在利用计算机编制《多语种语言学词汇（英、法、德、俄、汉）》的过程中没有造一个汉字，就是一个证明。但是，多少词才能象《基本集》那样满足人们的一般需要呢？北京师范大学中小学语言课本（12年制）的统计表明：106万8千字的语料中计有39601个词条。如果词库连中小学语文课本的词汇都包括不了，显然不能实用。

（2）从汉语词的结构考虑。前面已经谈过，现代汉语已不再是"单音节语"。单音节词只占百分之几（从静态统计来说），而绝大多数是多音节词，其中尤以双音节词为最多。划分汉语的词，必须考虑多方面的因素，即语法、语音、语义以及心理学、信息论等方面的因素。切分得太碎或过大，既不便于人的识读，也不利于信息处理。太碎，则同形词会增多，过大，则势必影响词库的单位储量（例如收了"国民经济"、"国民收入"、"国民教育"）等不少由2＋2组合的四字词语，就要出现这种情况①。考虑到汉语双音节化的矛盾②，认为二、三音节的词占绝大多数恐怕是汉语的现实。另外一部分四字格可作词或词组词。

（3）从"视读原则"考虑。为了信息处理的需要，把词的长度

①最近看到几本词表，有的词表以"经济"打头的四字词语收了三十多个，但还缺了"经济建设"、"经济效益"、"经济杠杆"词语、另一份词表收了一百多个，上面几个都收了进去。可以说够大够全了。然而却缺了第三本词表中收集的"经济小吃"、"经济掠夺"、"经济渗透"、"经济竞赛"、"经济战线"、"经济封锁"、"经济警察"、"经济杂交"、"经济观点"、"经济斗争"、经济昆虫"、"经济命脉"等十二个词语。有意思的是，第三本词表仅有二十八个这样的四字词语，而其中十二个竟然是那本最全的词表所没有的。因此，可以得出这样的结论：象这样收词，真是收不胜收。倒头来，词库容量数目上增加了，实质上并没增加，甚至会减少。

②参见周有光："正词法的内在矛盾"，《文字改革》，1983年第3期。

拉长一些是完全合理的，这也是拼音连写的需要。例如，将静态的平均词长1.98音节（语言学院《汉语词汇的统计与分析》）拉长到2.2—2.3音节，看来是合理的。但是，拉长的限度要以不违反心理学和信息论的视读原则为准。汉语音节由1至6个字母组成平均长度为3.01字母（文改会及杭州教育局的统计，见《文字改革》1985年第1期），声调如以字母标记，就要接近4。因此，平均词长为2.2音节，则意味着平均为6.5—8.5个字母。看来，这样的长度还在视读原则所允许的范围之内。如果考虑到动态的平均词长会降到1.7个音节左右，那么就可接近信息论的最佳值了。

（4）从信息处理角度考虑。对于信息处理来说，标准化很重要。什么是一个词，如果以拼音形式出现，则完全可以采用“两个空格之间的一组字母”这个简明的定义。为了便于人们掌握分词标准，似乎还可加一个限制，即“汉语的词不得超过六个音节”。我国翻译外国地名时，一般也是以六字为准，超过时加短横或省略轻读音（详见《外国地名译名手册》，1983年）。从信息处理角度来说，有必要区分构形成分（“们”、“了”、“着”、“过”等词尾）和构词成分（“非”、“超”、“化”、“性”、“主义”、“分子”等词缀）。一般来说，前者属于开放类，类似于“-s, -ed, -ing”之类的词尾，而后者属于封闭类，是本身的组成部分。这样，进行词汇统计时，可分别处理，前者可与其主体词分别统计，而后者则与其主体词当作一个整体统计。合理分化“的”、“地”、“得”对于机器翻译、人机对话等课题来说是大有好处的，即分化出作形容词词尾的“-d”，作结构助词的“de”作副词词尾的“-di”和作动补词尾的“-de”。另外，充分利用短横的作用也有助于提高识读效率和信息处理能力。

至于在此词库的基础上提供什么信息或特征，那是下一步要考虑的问题。不过，这里可以先说明一点，许多信息特征应根据专门目的而定，其中最基本的信息可说是词类信息，因为许多自

然语言理解系统都离不开它。但汉语词类目前还是一个悬而未决的问题，因此，又是一个需要我们克服的困难。语义信息也是必需的，但目前尚无成熟的经验，国内外学者正在探讨。

10.4 计算辞书学 编制目录、索引、词表和词典，对于人来说是相当繁琐的工作，但对电子计算机来说却不然。因为它最擅长统计、分类、编排、检索之类的工作，也不嫌浩繁的简单重复性劳动。电子计算机的应用使辞典业的发展出现了新的契机，一门崭新的计算辞书学正在兴起。

10.4.1 文献目录的编制 这种程序并不难编，它跟各种档案的程序、工资表程序大同小异。主要的问题是项目的安排。一般设有下列项目：文献名称，作者，译者，合作者，分类，期刊名称及卷期，出版单位，出版地，出版时间，国别，文种，另种文本，有无评论，是否被引用，备注。项目安排得合理，程序编制得好，从一份资料档中便可得出所需要的多种结果。例如，可得出作者目录，分类目录，译著目录，受到评论的文献目录，近十年来某一国家用英文发表的文献目录，等等。

10.4.2 索引的编制 索引种类很多，字顺索引，附参照的索引，引文索引，分类索引，组配索引，语汇索引，累积索引，关键词索引（包括关键词与题名索引 KWAC index，题内关键词索引 KWIC index，题外关键词索引 KWOC index），地址索引，轮排索引，期刊索引，等等。这里只谈语汇索引方面的问题。语汇索引（concordance），从语汇单位的长度来分，可分四种；从编排格式来分有三种。

(1)单字索引。这是我们使用汉字的人所特有的。为古代文献编过百十种，有叫引得的，也有叫通检的。武汉大学计划编制的现代文学作品索引33种，就属于这种类型（第一辑《骆驼祥子》索引已于1983年印出）。最近编成的《寒山子诗》逐字索引也属于这种类型。①

①见《中国语文》1985年第3期。

(2)单词索引。这是一种以词为单位的索引，现在国外大量编制这种索引，为语文研究和辞书编纂工作提供了极为有利的条件。有的把整部著作出现的所有词都作了索引，因而叫逐词索引，有的只为编者认为是重要的词作索引，因而叫重要词索引。

(3)词组索引，或叫短语索引。比方对“研究研究”“考虑考虑”“干干净净”“高高兴兴”这样的重叠结构作一番调查，或对各种词组、成语的应用进行调查，并制成索引。不过，在编制后一种索引时，常要有词典支持，即首先通过词典明确哪些词为词组。

(4)结构索引。这里的结构与上面谈的词组不同，它中间可插入其他词。比方可以指明让机器查“非……必”“非……不”“非……则”“非……而何”等结构，或查“虽然……但是”“不仅……而且”等结构，并制成索引。编制这种索引时，除需词典支持外，还要划定查找范围，一般是以句子性标点(句号，问号，感叹号)为界。

(5)地址型索引。这种索引只打印出该语汇单位出现的地址，如哪本书第几页第几行。好处是省纸，缺点是不直观。

(6)片段型索引。这种索引除突出地打印出该语汇单位(字、词或词组都可打在一行的中心位置)外，还将其左右的若干字词打印出来，打满该行为止(宽行打印机，一行可打132个字符)，而不管是否为一句。国外很多索引是这种类型的。好处是一目了然，缺点是句子时常不完整。

(7)整句型索引。这种索引的好处是能给出该语汇单位的完整的上下文。这样一来，该语汇单位的出现环境就变成不定长了。为了突出该语汇单位，西文可用斜体，中文可用不同字体或不同字号。当然也可以用不同颜色。这种索引对中文来说还有一个好处，就是省纸，因为每个汉字一般占1.5字符或2字符的位置，一行能打66或88个汉字，而汉语句子一般都不会占一行。

10.4.3 词表的编制 在言语统计的基础上可以得出多种词表，如正序词表，逆序词表，频度词表，按字母个数排列的词表，按不同领域编排的词表，等等。

（1）通用的最低限度词表和各专业的最低限度词表。这是最简单的一种词表，如不加注解，输入之后，机器可自动完成。一般在统计几十万到一百万词形（或字）的基础上便可得到比较可靠的各专业的最低限度词表（或字表），但是要想得到比较可靠的通用的最低限度词表（或字表）至少要统计几百万。北京语言学院出版社最近出版的《常用字和常用词》便属此类。

（2）词组和结构频度表。用上述词组索引和结构索引的相同点在于都需要词典支持，不同点有三：1）词表是由多种来源的资料归纳出来的，而索引多半是由一种作品或一个作家的作品得出的；2）词表一般只列频度，有的虽列分布面，但不列具体出处。而索引既重视频度又重视身体出处，有的还给出具体的上下文；3）词表一般供学习用，而索引则为研究用。

（3）逆序词表。对词尾变化多的语言来说，逆序词表价值最大。对于汉语来说，逆序也很重要。我们可以从逆序词表中得出如下信息：1）韵的所在（我国的《佩文韵府》就是这种类型的辞书），2）义的中心（如“彩旦”“花旦”“老旦”“刀马旦”……），3）语法的标志（“的”“地”“子”“性”“化”等词尾）。更重要的是：可以借助逆序提供的线索来区分同义（如“浮夸”“矜夸”“虚夸”），检查漏收的词条，以及研究构词方法及各词缀的能产性。

（4）双语或多语术语词表。在统计调查出各科术语的基础上，由人加上本族语的对等术语，输入计算机后，经过排序、查询等手续便可打印出按字顺排列的双语或多语词表，同时还可打印出另一种或多种索引。如果事先为某个词条按范畴进行分类，那么还可以打印出范畴表。中国社会科学院语言研究所正在编制的《多语对照语言学词汇》就属此类。人们可通过查英文正文或查一种文字的索引而得知英、法、德、俄、中五种文字的术语，也可通过查范畴表得知该术语在整个术语系统中的位置及其有关的术语。

10.4.4 词典的编制　有了各种索引和词表，加上丰富的语料库和知识库，编制词典便有了可靠的基础。利用计算机编词典，主要

是指材料的收集整理，资料的统计分析，索引词表的编制，语料库和知识库的建立和调用，词条的编制、修改和增删，词典版式的安排，输出纸版本和输出磁版本(软盘、磁带等)，另外还可缩微或激光照排印刷。不难看出，这些工作基本上是技术性的工作。至于词义的分析和注释，计算机无能为力，还只能由人来负担。需知，目前决定一部辞书的价值和实用性的，主要还是靠编辑人员的知识和技能。当然在编词典时，人们可以得到计算机很多帮助，如根据正序和逆序词表查重条、查漏条，根据频率特征注明常用的等极，根据大量方便的资料归纳出科学的词义，并给出适当的例句。甚至，为了词典的通俗性，还可以象《Longman 当代英语词典》那样，使得词典释义和举例中出现的每一个词都在编者规定的两千个词以内。为了做到这一点，计算机要进行大量比较对照工作，发现超出的词以后，由人从两千词以内选出适当的词加以代替。

10.4.5 综上所述，利用计算机编制辞书有下列优点：

(1)既快又好，可省去排版过程，同时也意味着少出错误和免去多次校对的劳苦。利用计算机编制索引尤为方便，可以省去繁重的人工倒片工作。

(2)可同时提供两种版本：纸版本和磁版本。

(3)修订方便，能紧跟潮流。可随时修改补充，不断更新版本。

(4)灵活多用。磁版本是一种活的工具，人们可以根据自己的要求来检查各种数据，例如可以查询某个国家的文献被译成外文的有多少，哪些著作被引用，受到评论。这些数据还可以用来作为评定优秀著作的部分凭证。

机编工具书的好处是十分明显的，因而这项技术越来越受到人们的重视，并为很多辞书界人士所采用。象 Webster 和 Oxford 这样大型的词典，经过长期酝酿，竟然也决定改用计算机来编纂，这足以说明，这项技术已达到相当成熟的地步，得到了社会公认。

但是，必须指出，对于使用汉字的人来说，还有一些特殊问题要考虑，如定序问题和分词问题。

条目如何编排，索引如何编制，不仅是人工编制各种工具书的重要问题，而且也是机编工具书的重要问题。目前常见的查字法有音序法、部首法、笔画法和四角法等几种。这几种查字法都可以让机器实现。但是，音序法实现起来最方便。目前，我们的音序法有两种：一种是纯字母排列法（如《中国大百科全书》），另一种是音序同字头归类排列法（如《现代汉语词典》）。前者更便于计算机处理，它是各国通用的排列法，因而很多计算机有根据纯字母排列原则自动排序的功能；而后者实际上还是一种未摆脱汉字束缚的排列法，实现起来有很多麻烦。如果想实现部首法、笔画法、四角法、或音序同字头归类排列法，那就需要在输入时为每个字增加"部首""笔画""四角号码""汉字代码"等信息。例如，武汉大学编制《现代汉语语言资料索引（第一辑骆驼祥子）》时，为了得出部首索引，分别为188个部首编定了由两个字母组成的代码，如氵(DS)、宀(BG)、扌(TS)、艹(CT)、口(KR)，等等；并给每个汉字加上了这样的代码，如"安"(AN_1NBG)、"案"(AN_4MBG)、"守"($SHOU_3CBG$)、"宣"($XUAN_1HBG$)，等等（注：数码为调号，数码前为汉语拼音，数码后第一个字母为区分同形的代码，最后两个字母为部首代码）。

由此可见，在推广汉语拼音应用的同时，大力推广纯字母排列法是十分必要的，起码是对中文信息处理极为有利。

10.5 普通话与电脑 普通话，即标准的中文，不仅有其书面形式，而且有口语形式。书面信息的处理已相当复杂，但比起语音信息处理来，则是小巫见大巫。

10.5.1 语音信息处理的原理及概况 这里所说的语音信息处理包括自动汉语识别和汉语合成两个方面。从目前科技发展水平来看，语音合成比语音识别取得的成就大，有些简易的合成器已投入实际应用。目前，合成方法有三：（1）单元编辑合成，（2）录

音编辑合成，（3）规则合成。第三种比前两种难度大得多，但有发展前途。

语音识别方面困难比较多，识别人的言语，光靠声学分析（语音模式匹配）不行，还必须进行大量的语言学分析。大致过程如下：语音模式匹配——→音节构成——→词的切分——→语法、语义、语用分析——→理解。如果是问答机，后一部分便是“回答”（即根据语音参数和规则合成，发出自然语言）。如果目的是进行翻译，在“理解”之后，还需进行两项工作：翻译和组成另一种语言（即根据另一种语言的参数和规则合成，发出另一种自然语言）。

从上面不难看出，“词”仍然是最基本的加工单位。只有切分出词，才能谈得上更高级的信息处理：自然语言理解和翻译。这种语音词的切分比起书面词的切分具有更大的复杂性，因为在语流中，词的要素或词本身受其周围环境的影响会发出各种变异（如上声和上声相连，则前一个上声变阳平）。

目前，识别器的适应功能不强，它对掌握同一方言的人所发的同一个音常常反应不同，误差很大。甚至对一个人在不同时间所发的同一个音的反应也有不同。比较好的识别器是“认人的”，即事先把该人的语音模式存储进去，然后再进行各种匹配。人们正在研究无特殊限制的识别器，总起来说，水平都还不够高，尚属试验研究阶段。

10.5.2 汉语方言的复杂性 下面来考察一下我们所要处理的对象——汉语。

汉语有众多方言土语。大的方言区有七个，各方言区还有数量不等的方言，方言之下又有一些次方言或土语。方言之间的差别有的较小，有的十分严重，达到人们彼此不能通话的地步。

汉语方言差别，主要在语音方面。从声调来说，不仅数目不同，而且调类除闽语区外也各异。如果再加上调值的不同，以及连续变调的差别，那就会构成一幅十分复杂的图景。如果拿语音全貌来讨论，那就不知还要复杂多少倍。

汉语方言的严重差别还表现在词汇方面。例如“母鸡”一词，在官话区的北京叫“母鸡”，成都叫“鸡母、鸡婆”；吴语区的苏州叫“雌鸡”；温州叫“草鸡、鸡娘”；湘语区的长沙叫“鸡婆子”；赣语区的南昌叫“鸡婆”；粤语区的广州叫“鸡乸（下过蛋）鸡项（未下过蛋）”；客家（梅县）话叫“鸡嫲（下过蛋）、鸡嬡（未下过蛋）”；闽南（厦门）话叫“鸡母（下过蛋）、鸡䴔（未下过蛋）”；闽北（福州）话叫“鸡母”。

语法方面的差别小，但也还有一些。例如：普通话“说不过他”这句话如果用赣语来表达，则成了“话不佢赢（说不 他 过）”；用吴语表达，则成了“话伊勿过（说他不过）”。吴语的表达法在普通话中虽不多见，但还有这种语序，而赣语的表达法则根本不合普通话的语序。又如，粤语副词状语的位置不合普通话的语序：“你行先”（状语在谓语后）等于普通话的“你先走”（状 语 在 谓 语前）。

10.5.3 电脑的要求 大家知道，电脑目前不能自己思维，而只能根据人们给它编制的程序进行工作。因此，电脑在加工语言信息时首要的一条是要求语言形式化。这里所说的形式，既包括语言单位本身的形式，又包括它的位置和结构关系；既包括语音、语法、语义的特征，又包括语用的特征。所谓形式化，就是一切从形式出发，而不是从意义出发。要知道，电脑只认识形式，不懂意义。当然，这绝对不是说要完全抛弃意义。人在分析语言时总是要参考意义的，因为许多场合完全靠形式行不通。但是，可以供电脑作依据的绝不是纯粹的意义，而是形式化了的意义。总之一句话，只有形式化，才能算法化、自动化。

随着形式化而来的要求主要有两个：一是精密化，二是标准化。前者要求用精确的语言表达，切忌模棱两可。后者要求唯一地代表同类事物，切忌纷纭驳杂。

后面这个要求，同我们这个话题有密切关系。究竟是让电脑同普通话这一种形式打交道好，还是让它同各个方言 打 交 道 好

呢？如果让电脑回答，它也会说“只同普通话一种形式打交道好”。为什么这样讲呢？因为只有这样，科学上才合理，经济上才合算。同时对加强民族团结才有利。汉字编码问题就是一个教训。本来有一种编码最好，但是由于各种原因统一不起来（我国大陆上有四百多种方案，目前在各机器上使用的也有几十种，台湾省也有不少种）。为了解决这个矛盾，制定了信息交换用的内部标准码，尽管增加了一道翻译手续，但毕竟是使外部互不一致的码在内部达到了统一。然而，这样的内部标准码，目前也还是有两套：大陆上一套，台湾省一套。这给信息加工造成了本来完全可以避免的麻烦，对信息共享来说无疑是一大障碍。如果推广普通话搞得不好，给电脑造成的麻烦必定会大得多，其影响也严重得多。

10.5.4 大力推广普通话是解决汉语语音信息处理问题的最好途径 问题非常明显：一方面是语言十分驳杂，另一方面是要求标准一致。这个矛盾如何解决呢？当然，随着研究的加深和条件的改善，机器的适应能力会越来越大。但总不能设想，机器能适应方言驳杂的情况。我们也不能设想，为每一个地区或每一个省都各搞一种识别器，不同地区之间要想通话再搞翻译器。事实上，如不立标准，为一个省搞识别器也没法搞，撇开语音不谈，例如湖北省就要把“媳妇”的22种说法都存入机器。显然，这样做，科学上不合理，经济上也不合算。唯一的出路是大力推广普通话，使语言的驳杂现象减少到最低限度（请注意不是要削弱普通话丰富多彩的表现力）。普通话是我们中华民族团结一致的基本保证之一。正因为如此，才把“国家推广全国通用的普通话”列入了我国的宪法。对推广普通话有利的事，人人都要去做，例如，编写一些方言区的人如何学习普通话的书；提醒作家在电影、话剧和文学作品中别滥用方言；制止为了赶时髦而乱用香港词语，等等。搞语音信息处理也要遵循这一点，因为这是唯一正确的方向。

为了更好地推广普通话，有必要加速普通话异读词的审定工作，尽可能消除同义异读现象，使普通话各方面的正音规范尽快

完善起来，并早日出版供广播员使用的正音词典(其中也包括外国人名地名读法)。

10.6 结语 中文信息处理包括内容很广，前几节中已专门论述过几方面的问题。这里结合文字改革中的四定(定形、定量、定序、定音)工作谈了一些。同时也表达了作者的一些观点，即:语文工作是关系国家民族能不能迅速发展的大事；语文信息的计算机处理正在世界范围内蓬勃开展起来；中文信息处理方面还存在不少问题(如“词”的问题，方音驳杂的问题)；大力推广普通话，努力扩大汉语拼音使用范围，并逐步过渡到汉字和拼音文字的双轨制，是解决问题的合理途径。

最后，希望语文工作者，尤其是文字改革工作者考虑中文信息处理的要求，以便早日建成现代化语言信息系统，使我国顺利地过渡到信息化社会。同时希望广大读者在了解了语文发展前景之后，努力使自己在语文修养方面符合信息化时代的要求。

六五三四
三甲

参 考 书 目

（1）草间基[日本]等著，侯汉清、肖自力译：《情报管理入门讲座》，科学出版社，1981。

（2）*Georg F. von Ostermann: "Manual of Foreign Languages"; New York, Central Book Company Inc.*, 1952.

（3）黄绳熙：《CWIS——一个中西文兼容的综合情报检索系统》，科学出版社，1986。

（4）*Ignatius G. Mattingly: "Speech Synthesis for Phonetic and Phonological Models"*，美国百科知识出版社，1977。

（5）肯尼思、卡兹纳著，黄长著、林书武译：《世界的语言》北京出版社，1980。

（6）吕叔湘：《汉语语法论文集》(增订本)，商务印书馆，1984。

（7）刘涌泉：《语言学现代化和计算机》，武汉大学出版社，1986。

（8）刘泽先：《科学名词和文字改革》，文字改革出版社，1958。

（9）*MSZ asociation: "Translation of cyrillic characters into latin characters"*, *KGST*-1362-78，1982。

（10）P. C. 吉利亚列夫斯基、B. C. 格列夫宁合编，杜松寿译：《世界各种文字样品》，文字改革出版社，1958。

（11）清华大学外语系：《清华大学外语系计算机辅助教学系统介绍》，清华大学出版社，1987。

（12）乔毅：《法汉题录机器翻译试验》，载《中文信息学报》，1987年第4期。

(13) *Wayne A. LEA:"Computer Recognition of Speech"*，美国百科知识出版社，1977。
(14) 杨沛霆：《科技情报工作的几个问题》，科学出版社，1981。
(15) 曾世英：《中国地名拼写法研究》，测绘出版社，1981。
(16) 中国地名委员会：《外国地名译名手册》，商务印书馆，1983。
(17) 朱川：《实验语音学基础》，华东师范大学出版社，1986。